Vibrational spectroscopic image analysis of biological material using multivariate curve resolution–alternating least squares (MCR-ALS)

Judith Felten[1], Hardy Hall[1], Joaquim Jaumot[2], Romà Tauler[2], Anna de Juan[3] & András Gorzsás[4]

[1]Umeå Plant Science Centre, Department of Forest Genetics and Plant Physiology, Swedish University of Agricultural Sciences, Umeå, Sweden. [2]Institut de Diagnòstic Ambiental i Estudis de l'Aigua–Consejo Superior de Investigaciones Científicas (IDAEA-CSIC), Barcelona, Spain. [3]Chemometrics Group, Department of Analytical Chemistry, Universitat de Barcelona, Barcelona, Spain. [4]Department of Chemistry, Umeå University, Umeå, Sweden. Correspondence should be addressed to A.G. (andras.gorzsas@chem.umu.se).

Published online 8 January 2015; doi:10.1038/nprot.2015.008

Raman and Fourier transform IR (FTIR) microspectroscopic images of biological material (tissue sections) contain detailed information about their chemical composition. The challenge lies in identifying changes in chemical composition, as well as locating and assigning these changes to different conditions (pathology, anatomy, environmental or genetic factors). Multivariate data analysis techniques are ideal for decrypting such information from the data. This protocol provides a user-friendly pipeline and graphical user interface (GUI) for data pre-processing and unmixing of pixel spectra into their contributing pure components by multivariate curve resolution–alternating least squares (MCR-ALS) analysis. The analysis considers the full spectral profile in order to identify the chemical compounds and to visualize their distribution across the sample to categorize chemically distinct areas. Results are rapidly achieved (usually <30–60 min per image), and they are easy to interpret and evaluate both in terms of chemistry and biology, making the method generally more powerful than principal component analysis (PCA) or heat maps of single-band intensities. In addition, chemical and biological evaluation of the results by means of reference matching and segmentation maps (based on *k*-means clustering) is possible.

INTRODUCTION

Imaging is an essential tool for biological studies that aim to understand gene function related to developmental processes and phenotypical outputs. By using biochemical, molecular biological and spectroscopic tools to augment imaging, sample anatomy and morphology can be correlated to molecular characteristics. The term hyperspectral imaging is used for the techniques that use spatially resolved spectroscopic information to create images. Vibrational microspectroscopic techniques, such as FTIR and Raman microspectroscopy, are particularly suited for hyperspectral imaging of biological materials, as they are fast, noninvasive, nondestructive and inexpensive. They are also very versatile, and they provide molecular-level information with little to no sample preparation and without staining[1]. In addition, they can be used without a priori knowledge of sample composition, in contrast to immunolocalization studies of targeted bio-polymers[2,3]. These advantages make FTIR and Raman microspectroscopy popular in a broad field of sciences, ranging from medicine (e.g., identifying abnormal (cancerous) tissue areas on the basis of their chemical composition[4–6]) to wood biotechnology (e.g., analyzing the chemical composition of different cell types or cell walls[7–9]).

Thus, the primary application areas benefiting from our protocol include the biological and medical sciences, in which changes in chemical composition need to be interpreted in a spatial context. This includes mapping the effects of disease or pathogens, genetic modifications, environmental factors, or inherent biochemical differences between cell or tissue types.

In plant sciences in particular, vibrational microspectroscopy contributed to identifying the role of genes involved in the wood biosynthetic machinery[10–13] by enabling the mapping of chemical compositional changes in transgenic plants and at different developmental stages. The major types of biopolymers in woody cell walls (cellulose, lignins, hemicelluloses and pectins) have characteristic FTIR and Raman spectroscopic fingerprints[8,14–22], making vibrational microspectroscopy suitable for the chemical imaging of wood. Historically, FTIR microspectroscopy has been more commonly used for this purpose[23], mainly because of the fluorescence problems encountered in Raman spectroscopy because of lignin. However, with the development of strategies to circumvent fluorescence problems, Raman spectroscopy is rapidly gaining popularity owing to its high spatial (confocal and lateral) resolution[8,12,14–16]. This high spatial resolution enables the chemical analysis of distinct (sub)micron-sized zones or layers within woody cell walls[8,14], providing an advantage over standard FTIR microspectroscopy, which usually aims at single-cell resolution only[7]. Plant cell walls are heterogeneous mixtures of mainly four major types of biopolymers: cellulose, hemicelluloses, lignins and pectins. The structure, composition and relative proportion of these biopolymers can vary not only between tissues and cell types but also along a developmental gradient and even within a single cell wall. Cell walls of woody tissues in plants are characterized by their layered structure, with different layers having distinct proportions of the above-mentioned polymers and distinct ultrastructural features. Although different cell wall layers may not be visible in white light images, they can in theory be differentiated on the basis of their spectral profiles: i.e., chemical composition. This is of great importance when the effects of, e.g., genetic modification or environmental signals on cell wall layer development and composition need to be investigated. In such studies, it is imperative to compare the same cell wall layers in wild-type and transgenic or untreated and treated plants, without the influence from the neighboring cell wall layers or cells. Raman imaging is capable of providing the necessary resolution, and we provide an example in the ANTICIPATED RESULTS section.

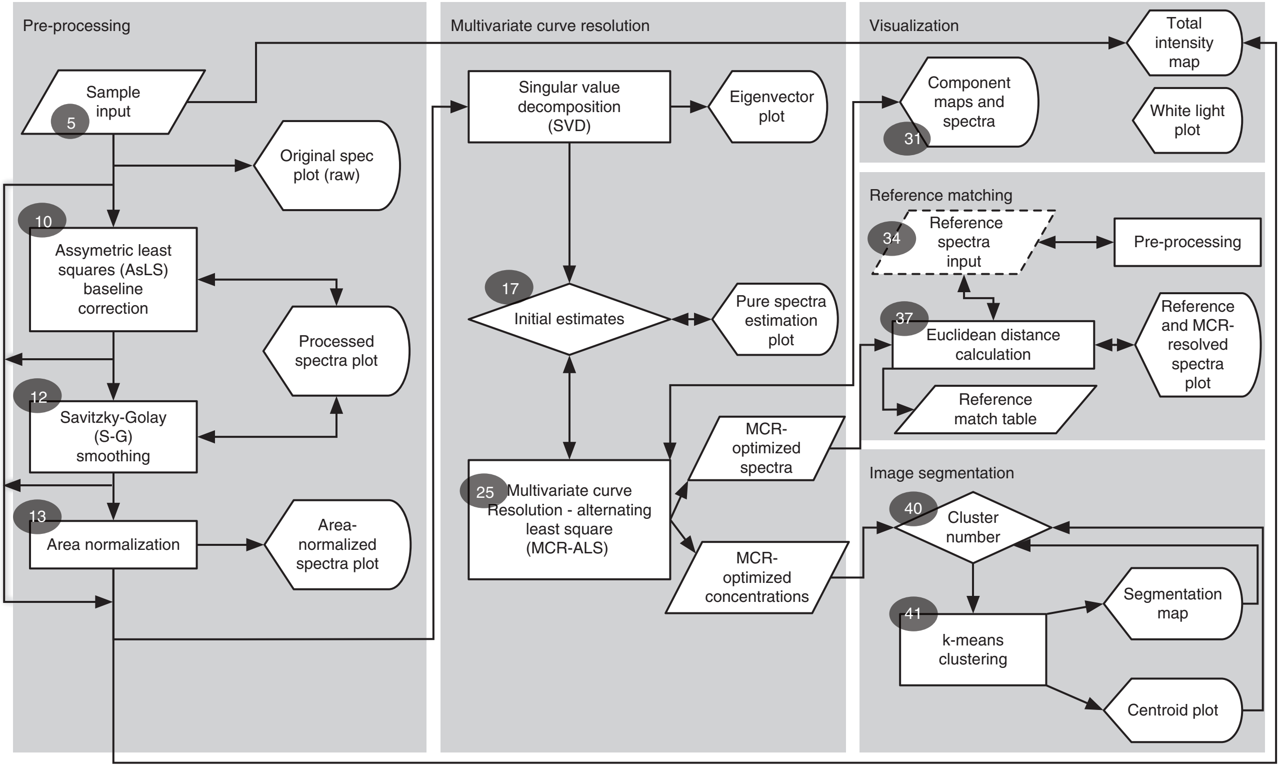
Pre-processing
Sample input
5
Original spec plot (raw)
10
Assymetric least squares (AsLS) baseline correction
Processed spectra plot
12
Savitzky-Golay (S-G) smoothing
13
Area normalization
Area-normalized spectra plot
Multivariate curve resolution
Singular value decomposition (SVD)
Eigenvector plot
17
Initial estimates
Pure spectra estimation plot
25
Multivariate curve Resolution - alternating least square (MCR-ALS)
MCR-optimized spectra
MCR-optimized concentrations
Visualization
Total intensity map
Component maps and spectra
31
White light plot
Reference matching
34
Reference spectra input
Pre-processing
37
Euclidean distance calculation
Reference and MCR-resolved spectra plot
Reference match table
Image segmentation
40
Cluster number
Segmentation map
41
k-means clustering
Centroid plot

Figure 1 | ** Overview flowchart listing the steps of the data analysis protocol. The data analysis can be broken down into five major parts (gray shaded areas, which also reflect the layout of the GUI (Supplementary Fig. 1**)): Pre-processing, Multivariate curve resolution, Visualization, Reference matching and Image segmentation, each having a number of steps and options. The major stages are labeled with gray ellipsoids and numbers that correspond to the Steps in the PROCEDURE section. Data analysis starts with Sample input and follows the arrows, although some steps are optional. For details, consult the running text. Parallelograms indicate data matrices, rectangles indicate processing steps, diamonds indicate key conditional choices and bubbles denote plots.

In addition, we illustrate the use of hyperspectral imaging in medical sciences, in which the challenge lies in the clear identification of chemical differences in order to assess pathological conditions with certainty. For instance, the monitoring of disease progression requires a set of quantitative markers that are clearly associated with (the different stages of) the disease and can be accurately followed. The ANTICIPATED RESULTS section contains an example using a mouse pancreas section, in which we identify the islets of Langerhans without staining, purely on the basis of spectral information. Monitoring the size and distribution of these islets provides a valuable tool for tracking the progression of type I diabetes and illustrates another application area of the protocol.

Vibrational microspectroscopic image analysis

The three main steps of gaining chemical information using vibrational microspectroscopic techniques are sample preparation, sample measurement and data analysis. Detailed protocols have been developed for plant biologists regarding sample preparation and spectroscopic measurements[7,14,24]. However, data analysis often stops at the level of generating simple band intensity maps (heat maps), or it is performed on a case-by-case basis by specialists in the fields of chemometrics and spectroscopy[7]. As a result, biologists are often disconnected from the data analysis steps, making them less likely to use vibrational microspectroscopy in their research.

Therefore, we present a pipeline (**Fig. 1**), together with a detailed protocol and software script (**Supplementary Fig. 1**) specifically designed for users with limited background in chemometrics and spectroscopy. This protocol is centered on MCR-ALS analysis streamlined for a well-defined kind of data analysis problem, namely vibrational microspectroscopic image analysis of biological samples. Accordingly, most MCR-ALS parameters are preselected, and certain analytical options are entirely omitted (see **Table 1** and the ‘Additional comments and limitations’ section of the INTRODUCTION) for ease of use and to make it possible for biologists to rapidly and reliably analyze their own data, and to gain a quick overview of a large number of samples. Consultation with specialists can thus be restricted to the most complicated cases, for which the basic application of the algorithm may be insufficient, or to the more advanced steps of the analysis (e.g., comparing the results of different models, spectral band interpretations).

The protocol integrates the most suitable procedures for advanced vibrational microspectroscopic image analysis into a complete package via an interactive (freely available as a MATLAB script at http://www.kbc.umu.se/vibrationaldownload.html). Together with the already existing protocols for sample preparation and recording[14,24], it forms a complete vibrational microspectroscopy suite for plant sciences.

Although we focus on Raman microspectroscopy and plant material, the protocol is not limited to these. It can be directly applied to different kinds of hyperspectral images (including near-infrared (NIR), UV-visible, Raman and FTIR microspectroscopy) of any kind of biological samples (including medical, **Fig. 2**),

TABLE 1 | Constraints for single-image and single-method data sets.

Constraint	Possible choice in standalone scripts[40]	Possible choice in the present script
AsLS baseline correction	No baseline correction is possible	Value: *P* value adjustable between 0.001 and 1; λ value adjustable between 1 and 1,000,000,000
Non-negativity for concentration profiles	Value: can be set to yes or no individually for each component Method: can be set to non-negative least squares, fast non-negative least squares and forced to zero	Value: fixed to yes for all components Method: fixed to using fast non-negative least squares
Non-negativity for spectral profiles	Value: can be set to yes or no individually for each component Method: can be set to non-negative least squares, fast non-negative least squares and forced to zero	Value: fixed to yes for all components Method: Fixed to using fast non-negative least squares
Unimodality for concentration profiles	Value: can be set to yes or no individually for each component Method: vertical, horizontal or average. Tolerance level can be adjusted	Value: fixed to no for all components
Unimodality for spectral profiles	Value: can be set to yes or no individually for each component Method: vertical, horizontal or average. Tolerance level can be adjusted	Value: fixed to no for all components
Closure for concentration profiles	Value: can be set to yes or no individually for each component Method: fixed to a single value or allowing changing values. Equality or nonequality (equal or less than a preselected value) can be chosen	Value: fixed to no for all components
Closure for spectral profiles	Value: can be set to yes or no individually for each component Method: fixed to a single value or allowing changing values. Equality or nonequality (equal or less than a preselected value) can be chosen	Value: fixed to spectra equal length
Equality for concentration profiles	Value: can be set to yes or no individually for each component Method: selectivity or local rank information via an auxiliary matrix. Equality and nonequality (equal or less than the predefined value in the auxiliary matrix) can be chosen for the constrained concentrations	Value: fixed to no for all components
Equality for spectral profiles	Value: can be set to yes or no individually for each component Method: known compound spectra via an auxiliary matrix. Equality and nonequality (equal or less than the predefined value in the auxiliary matrix) can be chosen for the constrained spectra	Value: fixed to no for all components

For detailed descriptions of the constraints, refer to the INTRODUCTION and Jaumot *et al.*[40] and references therein.

and it covers all the important stages of the analysis in several steps: (i) pre-processing of the spectra; (ii) MCR-ALS analysis of the data, focusing on unmixing the different chemical constituents; (iii) evaluation of the results in terms of chemistry by reference spectra matching; and (iv) visualization and evaluation of the results in terms of anatomy using pure component maps and segmentation maps.

The main goal of this paper is to provide a step-by-step manual for each of these stages in the PROCEDURE section. In that section, we only provide information that is needed for practical decision-making and troubleshooting. The underlying principles, theory and other background information regarding all major parts of the analysis are detailed in the sections below.

Spectra pre-processing

Pre-processing of the recorded raw data is required before data analysis in order to remove all variation that is uncorrelated to, and interferes with, the chemical information in the spectra (fluorescence, background or total signal intensity variations, noise and so on). Below, we describe a procedure that accomplishes this task through baseline correction by asymmetric least squares (AsLS) fitting and optional smoothing and area normalization. This procedure assumes that the input data have already been corrected for basic method-related artifacts (e.g., cosmic rays) in accordance with previously described Raman imaging procedures[14].

Baseline correction. The most commonly required pre-processing step is baseline correction. In addition to environmental and instrumental sources (e.g., temperature or source intensity fluctuations, vibrations and so on), baseline variations can be caused by the inherent optical and physicochemical properties of the sample (edge effects, hot spots, autofluorescence, refractive index heterogeneities and so on). Although entirely linear baselines can theoretically exist, they are practically non-existent in real images. This limits the usefulness of simple one-point (offset) or two-point linear baseline corrections, especially in the case of Raman images of plants wherein there are additional contributions from fluorescence. Multipoint linear or polynomial baseline corrections can approximate real baselines better, but they can be

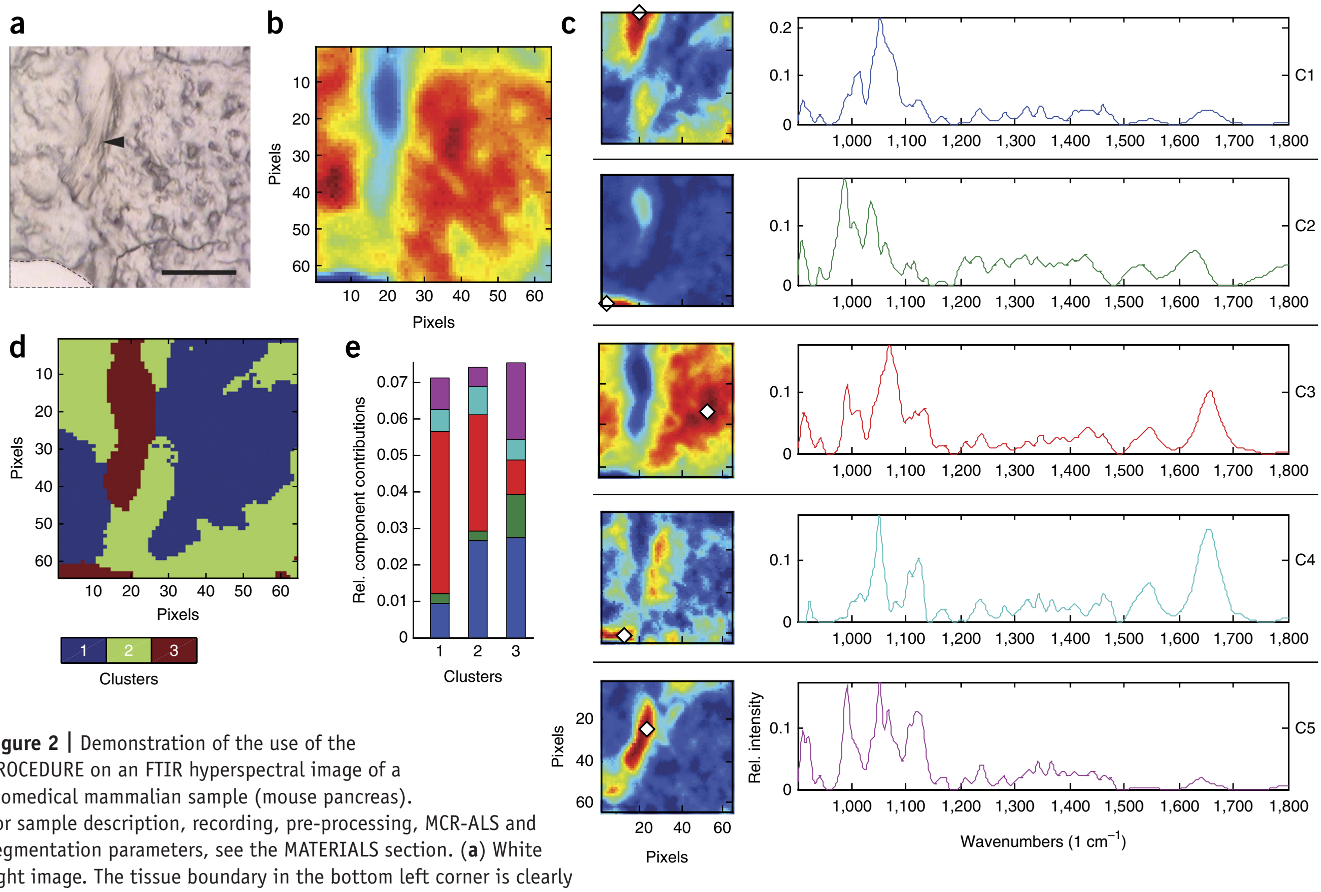

Figure 2 | Demonstration of the use of the PROCEDURE on an FTIR hyperspectral image of a biomedical mammalian sample (mouse pancreas). For sample description, recording, pre-processing, MCR-ALS and segmentation parameters, see the MATERIALS section. (**a**) White light image. The tissue boundary in the bottom left corner is clearly visible (area marked by dashed lines). Scale bar, 50 µm. (**b**) Total intensity map of the area shown in **a**, following pre-processing. The tissue boundaries are more visible (low-intensity regions shown in blue) and indicate a crack in the tissue section, as a result of drying. When comparing this heat map with the white light image, a region in the sample crack can be noticed, in which the embedding material is clearly visible (black arrowhead in **a**). (**c**) Pure component (C1–C5) distribution maps (concentration profiles, left) and the corresponding spectral profiles (right), following MCR-ALS. The white diamonds mark the location of the purest pixel for each component. On the basis of the spectral profiles and comparing the distribution maps of each component, it is clear that C2 and C5 (and to a lesser extent C1) contain contributions from the embedding medium. (**d**,**e**) Segmentation map and the corresponding centroid profile, respectively, following k-means clustering. The analysis is clearly able to differentiate distinct zones in the image. Cluster 1 (dark blue) corresponds to the islets of Langerhans, having a large contribution from C3 (red). Cluster 2 (light green) corresponds to cells of the exocrine tissue, with an increased contribution from C1 (blue) and C4 (cyan). Finally, cluster 3 mostly corresponds to sample-free areas of the image, with high contribution from signals of the embedding medium.

extremely difficult to perform correctly and reproducibly[25]. This is especially true in the case of vibrational spectra of biological materials, in which broad, overlapping peaks (or ‘bands’ in vibrational spectroscopic terminology) cover substantial areas of the spectrum and in which distinct image regions or pixels can have different, intense and irregularly shaped baselines. These features make it difficult to determine a fixed set of baseline points for polynomial fitting.

The method of choice for baseline calculation of vibrational microspectroscopic images of biological material uses AsLS fitting, originally proposed for chromatograms by Eilers[26]. AsLS baseline correction is an iterative method based on the use of a Whittaker smoother to fit the baseline to the data. It is fast, flexible, easy to perform and automate, and it only requires the selection of two parameters.

The first parameter is the λ value, which determines how smoothly (closely) the baseline is fitted to the data. Higher λ values result in a more linear baseline, whereas lower λ values generate baselines that closely follow the curvilinear natural baseline shape. λ values should be adjusted so that the corrected spectrum does not contain any broad features from leftover baseline, while at the same time it retains even low-intensity bands. Overly high λ values result in underfitted baselines, where the broad baseline features are not removed. Conversely, overly low λ values result in overfitted baselines, where spectral band intensities diminish to such extent that low-intensity bands could disappear entirely or suffer an artificial loss of intensity and distortion in the band shape (**Fig. 3**). λ values should always be adjusted to the working data set and tested on representative spectra of the image, by visually inspecting the fitted baseline and the resulting corrected spectra.

The second parameter to be adjusted is the *P* value. It gives a different weight to those points in the fitted baseline that have positive residuals (i.e., where bands or spectral features are present) as compared with those points that have negative residuals in the fitted baseline. As vibrational microspectroscopic images of biological material should not contain negative peaks, *P* values should be kept at the minimum value by default. It is important to

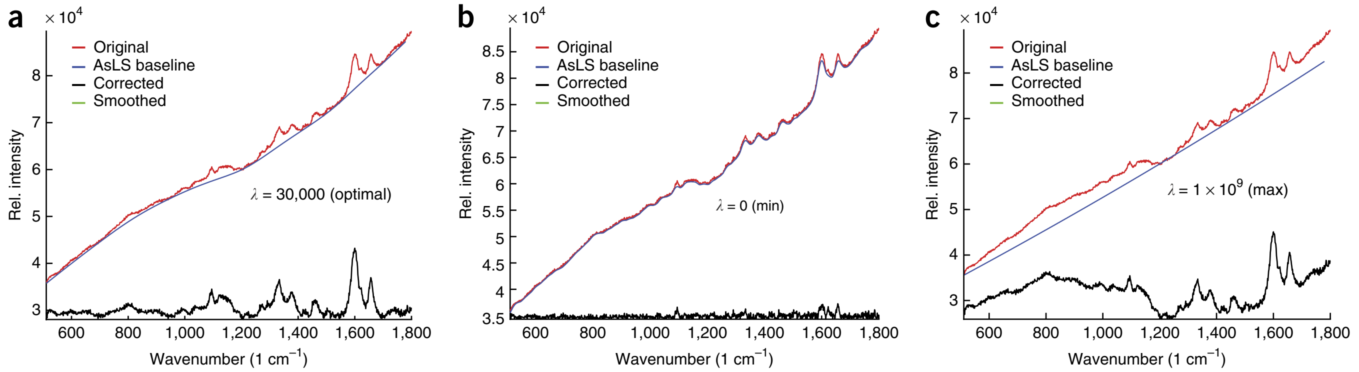

Figure 3 | The effect of different λ values on the results of asymmetrical least squares (AsLS) baseline correction. (**a**–**c**) The Raman spectrum of a poplar fiber cell wall is used as an example, showing the original spectrum in red, and the calculated baseline in blue. (**a**) The optimal λ value results in a spectrum that does not contain any broad features of leftover baseline, yet retains even small-intensity bands (black). (**b**) Overly low λ values generate baselines that follow the data too closely, resulting in overfitting, and as a result spectral band intensities diminish or disappear entirely (black). (**c**) Overly high λ values generate a very linear baseline, resulting in underfitting, and as a result the broad baseline features are not removed (black). The *P* value was kept at the value of 0.001 (default GUI minimum) for all plots.

note that this method is based on local fitting and does not apply any predefined baseline shape. This is why it adapts extremely well to data with irregular baselines, even if each spectrum of the data set has differently shaped baselines and intensities.

Smoothing. Smoothing is an optional pretreatment step that needs to be used with caution. It can improve data quality by removing noise, but excessive smoothing causes information loss by decreasing spectral resolution, which is easily noticed by the resulting distortion and the merging of bands (**Fig. 4**). Of the different spectral smoothing methods, Savitzky-Golay (S-G) filtering[27] is the most widely used, because it can improve the signal-to-noise ratio without distorting the signal. The smoothing is controlled by two factors: the polynomial order and the frame size (or number of convolution coefficients). In practical terms, the larger the difference between polynomial order and frame size, the more the smoothing applied. If the difference is 1, no smoothing is produced[28]. Smoothing parameters must always be visually evaluated from the graphical output on a case-by-case basis, to ensure that they result in noise reduction without signal distortion (**Fig. 4**).

Area normalization. Area normalization of the raw image spectra is entirely optional, as it is not required to describe the variation among pixels by multivariate analysis (**Supplementary Fig. 2**). On the contrary, area normalization is generally not recommended in resolution image analysis. However, it can facilitate the interpretation of component distribution maps in cases in which substantial intensity fluctuations exist (owing to the optophysical properties of the sample, such as pixel coverage, sample thickness, refractive index variations, out-of-focus effects and so on). It ensures that all pixels have the same total intensity, and thus the observed intensity variations in component maps represent (proportional) chemical concentration changes. In short, area normalization is only needed if substantial optophysical contributions are present, seriously hindering data evaluation or interpretation. In most cases, however, it will only affect data interpretation (in terms of amounts or proportions), but not the outcome of the analysis (**Supplementary Fig. 2**).

Data analysis

The aim of data analysis is to identify the chemical composition of the sample in a spatially resolved manner. This often includes the identification of chemically distinct zones and their correlation with biological information, such as anatomy, gene and/or protein expression, hormone gradients, enzyme activity and so on. Different approaches can be taken to identify chemically different zones. Selection can simply be based on anatomical features observed in the white light image (cell types, developmental stage and so on). Spectra from the selected areas or pixels can then be exported and the chemical information can be evaluated by different kinds of multivariate analysis[7,29]. However, visual identification of such zones can be problematic. First, there may be no visible features to base the selection on. For instance, the different

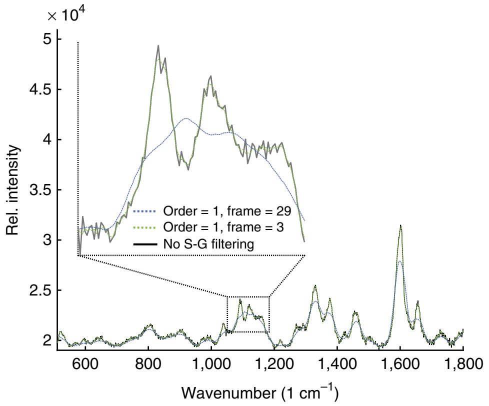

Figure 4 | The effect of S-G smoothing. The Raman spectrum of a poplar fiber cell wall is used as an example. The original (raw) spectrum without S-G smoothing is shown in solid black lines. The use of a mild smoothing (dashed green line, first-order polynomial, frame size = 3) reduces noise while at the same time retains band intensity, position and shape. The dashed blue line shows the result of oversmoothing (first-order polynomial, frame size 29). Although noise is undoubtedly reduced, band intensity, position and shape are all compromised. The inset shows a magnified spectral region for clarity.

cell wall layers of woody material cannot be distinguished in the white light image with the commonly used setups for vibrational microspectroscopy. Second, the chemically distinct zones may not (yet) be correlated to features visible in the white light image. For instance, at the starting phase of an infection process, healthy and newly infected cells may appear morphologically identical, whereas their chemical composition may already differ. In addition, finally, large sets of spectra from different samples may need to be compared, which requires a reliable (semi)automated and objective approach. This is the case when comparisons of genotypes, cell types or treatments need to be performed with many biological and technical replicates. Therefore, different methods have been developed to identify chemically different zones on the basis of the spectral information of the hyperspectral images in addition to the white light image.

Band intensity heat maps. The quickest, easiest, and thus the most commonly applied strategy for vibrational microspectroscopic data analysis is the mapping of integrated band areas (intensity heat maps)[8,18,30]. This method assumes the following: (i) that the nature of the chemical compositional change is known in advance (i.e., a priori knowledge of what to monitor); (ii) that this change only manifests in the intensity change of a well-defined spectral band and that this intensity change is only dependent on the (relative) concentration of the compound(s) associated with that band; and (iii) that the well-defined spectral band used for intensity mapping needs to be diagnostic of a particular compound.

The second assumption is invalid when bands overlap with those of another compound or shift in response to a chemical change, and thus the analysis will give erroneous results. Being dependent on this assumption also means that the data are sensitive to baseline correction errors and spectral artifacts (dispersive line shapes, noise and so on). The third assumption can be generally dangerous, as it does not consider the presence of unknown or unexpected compounds in the sample. Therefore, more reliable information is achieved by using a set of bands, or preferably the entire spectral profile, as not all bands are diagnostic.

Even in the case of perfectly diagnostic bands, from a single heat map alone no information is obtained regarding the relationship of the mapped compound to other compounds. Such relationships could better describe the chemical variation over the sample than the distribution of a single compound (**Supplementary Fig. 3**).

Heat maps of individual bands often show gradients within the sample, but no sharp boundaries. Cluster analysis based on the generated heat maps can be used to obtain segmentation maps and the corresponding centroid profiles, which help defining sharp boundaries and detecting ratio changes of different compounds. However, as segmentation maps are based on single-band intensities and not on full spectral profiles, they can still give erroneous results (**Supplementary Fig. 3**).

Finally, even if single-intensity maps and clustering provide a correct result, the analysis carries the risk of missing information, because only certain bands (partial information) are used. In other words, the analysis is significantly biased and unexpected changes will go unnoticed, because only a number of bands and only their intensities are monitored (broadening and shifts are ignored or interpreted as intensity changes).

PCA. PCA is a multivariate analysis method that takes the entire spectrum of the pixel into account to describe the variation within the image using a small number of basic contributions (principal components) related to spectral and spatial behavior[31]. It overcomes the above-mentioned problems of band intensity heat maps and is better suited to handle biological variation. Thus, PCA (alongside segmentation or cluster analysis) has gained popularity in the analysis of vibrational microspectroscopic images of biological materials[14,32,33]. However, interpreting the results of PCA in terms of chemistry can be difficult, as the scores (abstract distribution maps, **Supplementary Fig. 3b**) and the loadings (abstract spectra, **Supplementary Fig. 3d**) of principal components are estimated to be completely uncorrelated to each other, which is not the case for distribution maps and spectra of real chemical compounds. As a consequence, individual principal components cannot be generally associated with single chemical compounds or image regions (**Supplementary Fig. 3b**), and they may even contain information unrelated to the chemical compounds (e.g., scattering, fluorescence and so on) if suboptimal pre-processing has been carried out. Even in the case of optimal pre-processing, interpreting individual scores and loadings in terms of distribution maps and spectra of specific chemical compounds, respectively, is not straightforward because of the lack of one-to-one correspondence between chemical contributions and principal components (**Supplementary Fig. 3d**). In addition, several principal components may be required to differentiate distinct chemical zones. This further complicates data interpretation and visualization (**Supplementary Fig. 3h**), and makes it problematic to uniformly differentiate the same kinds of zones in several images. Finally, PCA loadings are generally hard to match to reference spectra, as they almost never represent the contribution of a single chemical compound, but a combination of spectral features originating from several compounds. Nevertheless, PCA is an important tool in multivariate analysis, and it is often the first multivariate method applied to a new data set.

MCR-ALS analysis. Another approach that shares the underlying bilinear mathematical model of PCA but considerably improves the interpretation of the results is MCR-ALS, followed by cluster analysis[34] (compare **Supplementary Figs. 2** and **3**). Essentially, MCR-ALS assumes that the complex spectrum in every pixel of an image can be described as a linear combination of the signal of a set of pure component spectra, conveniently weighted according to their abundance in each pixel. A 'pure component' in the context of MCR image analysis can be a pure chemical compound, or a part of the sample with a consistent spectral signature (i.e., a homogeneous mixture of compounds, in which case 'pure' means that it cannot be unmixed (resolved) further), and it does not have to be known before the analysis. The program is able to find the spectral signature (and the corresponding concentration profile) of each pure component, using only the image data (i.e., the spectra in each pixel of the image). The only required input is the number of components, which can be known a priori, or it can be determined using the variation in the image data (see the 'Determination of the number of components in the image data set' section below). In general MCR-ALS practice, pure spectral and/or concentration profiles can also be manually supplied to constrain the model. However, the present script does not allow for direct input of concentration or spectra of pure components (see the 'Additional comments and limitations' section).

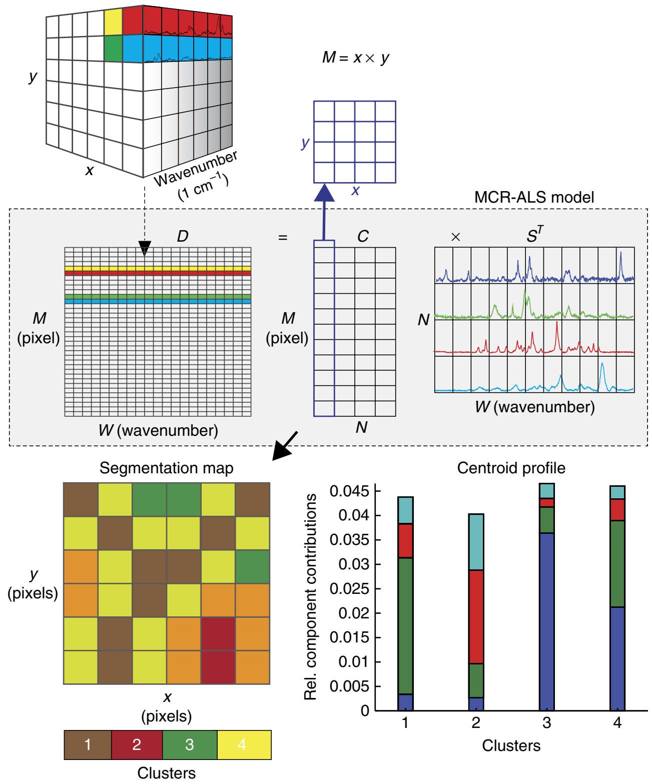

Figure 5 | Schematic illustration of multivariate curve resolution-alternating least squares (MCR-ALS) analysis on a hypothetical example. The hyperspectral image is described as a 3D data cube, with $x \times y = M$ spatial data points (number of image pixels) and W spectral data points (wavenumbers). The image is first unfolded (dashed arrow) to form the matrix D, which is the input for MCR-ALS. D is unmixed by MCR-ALS, using N number of pure components ($n = 4$ is used as an example in the figure), resulting in two matrices: C and S^T. C is an $M \times N$ matrix, every column of which contains the concentration profile of a pure component (the relative concentrations of the component in each pixel). S^T is an N × W matrix, every row of which contains the spectra of a pure component (illustrated as blue, green, red and cyan lines). As C is a reduced, noise-free representation of the original data (D), it can be used to construct segmentation maps and the corresponding centroid profiles. The segmentation maps use K number of clusters to describe K chemically unique zones in the sample ($K = 4$ is used as an example in the figure). Each cluster has different contributions from each pure component, as represented by the centroid profiles. The colors of the bars are the same blue, green, red and cyan, as used for the pure components for easy identification. For example, cluster 1 (brown in the segmentation map) has the highest contribution from the pure component represented by the green spectrum. Similarly, cluster 3 (green in the segmentation map) has the highest contribution from the 'blue' pure component and so on.

Under these conditions, MCR works by unmixing the original complex measurement (image data set) into the contributions of each of the pure components providing a signal. This makes it particularly suitable for the analysis of hyperspectral images. In mathematical terms, the MCR-ALS model is described as $D = CS^T + E$, where D is a matrix containing the spectra of all pixels of the image, S^T and C are matrices of the pure spectral signatures and the related distribution maps (concentration profiles) of the image constituents, respectively (**Fig. 5**), and E is the matrix of experimental error. Each one of the resolved (unmixed) pure contributions is thus represented by a pure spectral signature (a row in S^T) and a related distribution map, showing its abundance in each pixel of the image (a column in C).

The main difference with respect to PCA is that, instead of imposing orthogonality (i.e. lack of correlation) among components, MCR models the shapes of distribution maps and spectra according to natural chemical, spectroscopic and mathematical properties of the image data. For example, the concentration of a chemical compound is naturally non-negative. Likewise, normal, nonderivative spectra contain non-negative values only (baseline and positive peaks). Thus, concentration and spectrum non-negativity are natural properties of the data set. Optimization of the pure component spectra and the distribution maps is iterative until convergence is achieved.

There are other unmixing methods, e.g., vertex component analysis[35] or other algorithms coming from remote sensing. Their aim is also the linear unmixing of the raw image into chemically meaningful components[36,37]. Many of those are based on the calculation of a simplex that encloses all pixel spectra, the vertices of which are the endmembers (pure compound profiles) of the sample. Although these approaches are similar in aim to MCR-ALS and often provide comparable results, we preferred this iterative least squares–based algorithm because of a major freedom in terms of selecting and imposing constraints, the non-assumption of closure (which we do not consider a general applicable condition in image analysis[38]) and the simpler mathematical background compared with simplex-based unmixing approaches. A more detailed comparison of the performance of these unmixing algorithms is outside the scope of the present work, but it is available in the literature[39].

It is important to note that the results of MCR-ALS are not limited to the straightforward interpretation of pure component distribution maps and spectra. The maps provided by MCR-ALS can be used as starting (input) information for subsequent cluster analysis, as these maps are compressed representations of the information in the original image data set (**Figs. 1** and **5**). MCR followed by clustering is a powerful tool for identifying chemically distinct zones of the biological sample, either on the basis of pure chemical compounds or mixtures of pure compounds in similar proportions. It works well even for small hyperspectral image maps (**Fig. 6**), which is of great advantage when the scanned tissue zones need to be kept small to enable screening many samples in reasonable time.

Different software packages and detailed descriptions are available for applying MCR-ALS (and subsequent clustering)[34,40]. These are extremely powerful and versatile, but the variety of options and settings requires experience in chemometrics or spectroscopy for proper use. In particular, MCR-ALS optimizes the spectral and concentration profiles (distribution maps) of the image components under certain constraints. Constraints are criteria of biochemical or mathematical origin that the sought profiles must fulfill. We can divide the large variety of constraints into two main categories: those that are applicable for images and those that are designed for other kinds of data sets (e.g., processes). The latter includes constraints such as unimodality, closure, hard modeling and so on[41,42]. As these do not apply to the present protocol, they will not be addressed here any further. Among the constraints that are applicable for image analysis, the most commonly used is non-negativity. Concentration values in the distribution maps naturally fulfill this criterion, and pure spectra of compounds in vibrational spectroscopy also

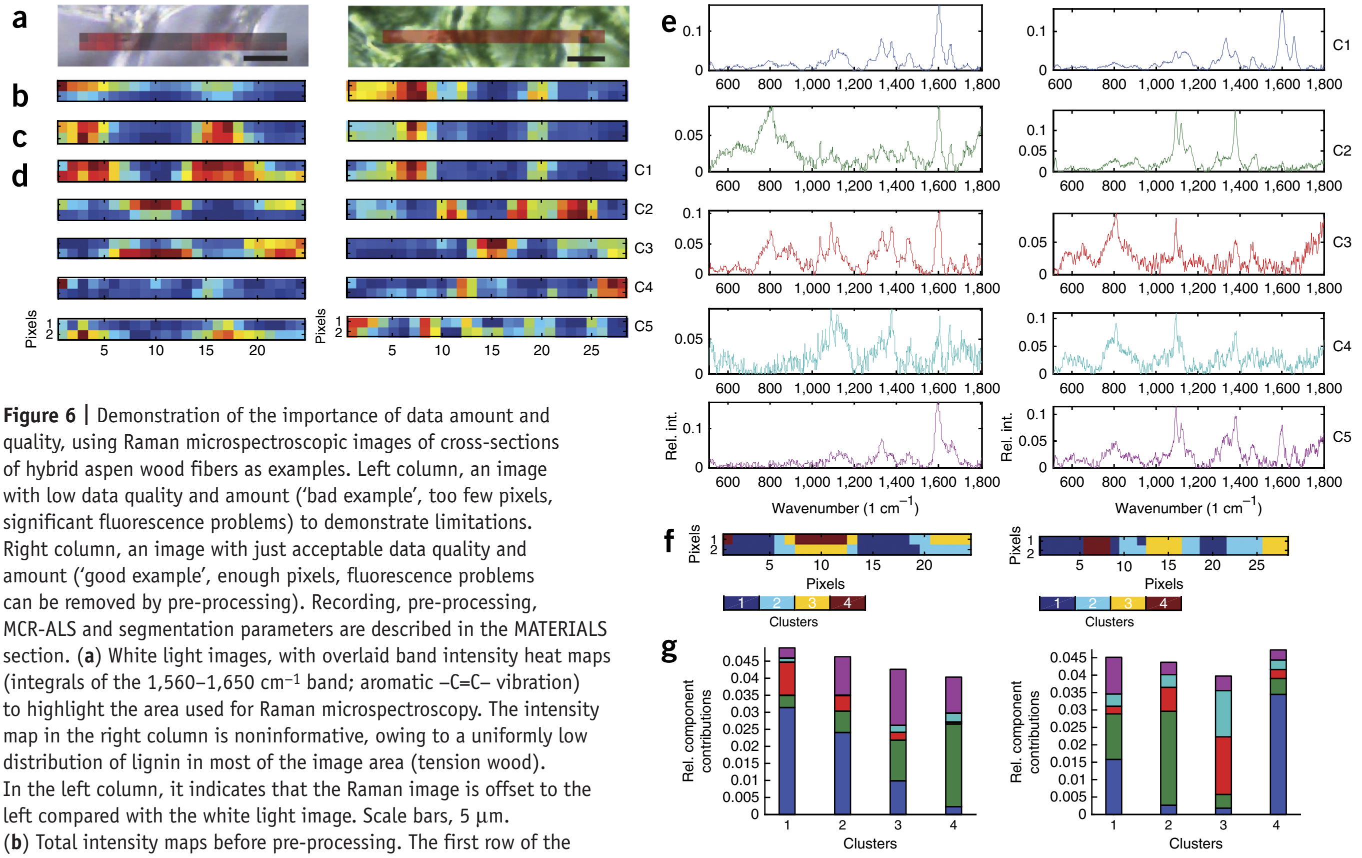

Figure 6 | Demonstration of the importance of data amount and quality, using Raman microspectroscopic images of cross-sections of hybrid aspen wood fibers as examples. Left column, an image with low data quality and amount ('bad example', too few pixels, significant fluorescence problems) to demonstrate limitations. Right column, an image with just acceptable data quality and amount ('good example', enough pixels, fluorescence problems can be removed by pre-processing). Recording, pre-processing, MCR-ALS and segmentation parameters are described in the MATERIALS section. (**a**) White light images, with overlaid band intensity heat maps (integrals of the 1,560–1,650 cm^{-1} band; aromatic –C=C– vibration) to highlight the area used for Raman microspectroscopy. The intensity map in the right column is noninformative, owing to a uniformly low distribution of lignin in most of the image area (tension wood). In the left column, it indicates that the Raman image is offset to the left compared with the white light image. Scale bars, 5 µm. (**b**) Total intensity maps before pre-processing. The first row of the bad example image has much higher (gradually decreasing) intensity, owing to fluorescence problems. The corresponding map of the good example image is free from this distortion. (**c**) Total intensity maps after pre-processing. The quality is considerably improved, matching anatomical features in the white light image (**a**). (**d**) Pure component distribution maps after MCR-ALS. Colors represent relative concentrations of the component from blue (low) to red (high). The models in the left and right columns are independent, and thus components do not have to match (see spectra in **e**). Maps on the left (bad example) only indicate the presence or absence of cell walls (compare with C1 and C2/C3). Moreover, the first and second rows in these maps are still visibly different, indicating that the MCR-ALS results still contain contribution from fluorescence artifacts. In contrast, the good example maps clearly indicate resolved cell wall layers (compare with C1 and C2). (**e**) The corresponding pure component spectral profiles. In the bad example, most component spectra have high noise level (except C1, corresponding to the sample area in the image, see **d** above). The lignin band (~1,600 cm^{-1}) is detectable in most of the component spectra. In addition, the spectra of C2 and C3 have high-intensity bands ~800 cm^{-1}, which is unrelated to the sample and coincides with the lumen (see **d** above). In the good example, the spectral profiles of C1 and C2 can be clearly associated with lignin and cellulose, respectively, whereas C5 is a mixture of both. The spectra of C3 and C4 have much worse signal-to-noise ratios and contain bands unrelated to the sample (~800 cm^{-1}). (**f**,**g**) Segmentation maps and the corresponding centroid profiles, respectively, using four clusters. For the bad example, cluster 1 describes the sample, whereas the other three clusters only describe different parts of the lumen, with more or less contribution from fluorescence or stray signal close to the sample boundaries (cluster 2). It is also confirmed by the centroid profiles (**g**). For the good example, the clusters can be associated with distinct cell wall layers. Cluster 1 is the secondary cell wall, with a high contribution from both cellulose and lignin (C2, green and C1, blue, respectively). Cluster 2 is the G-layer, having the highest contribution of cellulose (C2, green). Cluster 4 is the middle lamella, having by far the largest contribution from lignin (C1, blue). Cluster 3 is the lumen, with high contributions from components unrelated to the sample (C3, red and C4, cyan).

contain zero (baseline) or positive values (bands) only, unless specific operations (spectral subtraction or derivation) have been performed. To avoid intensity fluctuations in the recovered pure spectral signatures, the pure spectra in S^T are also normalized via a spectral equal length constraint[40]. These basic constraints can be applied to all vibrational image data sets and are prefixed in the present script for ease of use.

It is important to note that MCR-ALS is an iterative method, and as such the correct progress of the optimization depends on several aspects. The first is the starting point for the optimization, which is determined by the number of components and initial pure spectral estimates. The second is the endpoint of the optimization, which is determined by the number of iterations and the convergence limit and by the subsequent quality assessment of the results obtained. We provide guidelines for selecting the correct values for these factors, together with notes on how to evaluate whether the optimum results have been obtained by MCR-ALS or whether the analysis needs to be repeated to test a different set of parameters.

Determination of the number of components in the image data set. The number of components must be determined before MCR-ALS can be applied. In some cases, this may be known already beforehand (e.g., mixtures with known components). However, when no a priori information is available, singular value decomposition (SVD) can be used to estimate the number of components. SVD is an algorithm that describes the data set by an abstract model of bilinear contributions, formally analogous to the MCR-ALS model. Assuming correct pre-processing (i.e., elimination of intense baseline contributions or other artifacts), the relevance of each component is defined by the magnitude of its singular value: high singular values relate to biochemical contributions and small values to noise. The number of components that have high (clearly different from noise-related) singular values is therefore chosen as the total number of components used in the MCR-ALS model. However, in biological images, the exact number of components is often not evident from SVD values alone (**Fig. 7**). It is therefore recommended to test several models using different numbers of components and to evaluate the results in terms of model fit quality and interpretability of the final maps and spectra obtained.

Initial estimates. To start the optimization process by alternating least squares, an initial estimate of the pure spectral signatures (the matrix S^T) should be available. In the context of image analysis, the best option is to use a method that is based on the selection of the purest spectra in the image data set. For the present work, we chose a method based on SIMPLISMA (SIMPLe-to-use Inter- active Self-modeling Mixture Analysis)[43] to determine the initial pure spectra estimates. SIMPLISMA is a method based on the sequential choice of pixels with the purest spectral signatures within the raw image data set. It works well even on hyperspectral images of any size, and it has been used successfully in various Raman and FTIR microspectroscopic image analyses[43–45].

Evaluation of the results

The main criteria for the quality assessment of the MCR-ALS results are satisfactory model fit (mathematical evaluation) and meaningful distribution maps and spectra (biological, chemical and spectroscopic evaluation). We provide the basic information for evaluating each of these criteria. It has to be noted that no priority among these criteria can be set, and thus they should always be evaluated together.

Input data visualization (spectroscopic evaluation). Although chronologically this step precedes MCR-ALS (see PROCEDURE), it is part of the evaluation tools. Total (non-normalized) intensity plots that are generated before and after pre-processing can be used to determine whether the chemical image overlaps perfectly with the white light image or whether there are other distortions (out-of-focus effects, fluorescence problems and so on) that could affect component mapping and segmentation. It is an important tool for ensuring that data quality after pre-processing is good (i.e., artifacts are removed), and that signal intensities match anatomical features (**Fig. 6**).

Model fit (mathematical validation). The lack of fit at the end of the MCR-ALS process is the first parameter to indicate whether the model describes the data well. A satisfactory result should be reached with the selected number of components; i.e., the uncertainty of the model (lack of fit) should be in agreement with the experimental noise in the data. There is no default optimum value, as it strongly depends on the quality of the original data. When the lack of fit is unexpectedly high, several options are available. The best general strategy is to test models with an increased number of components and to see how this affects the model fit. If the fit improves, the original model did not include sufficient information to describe the system. If the model fit worsens or does not improve significantly after adding a new component, it is likely to be an unnecessary compound, unless

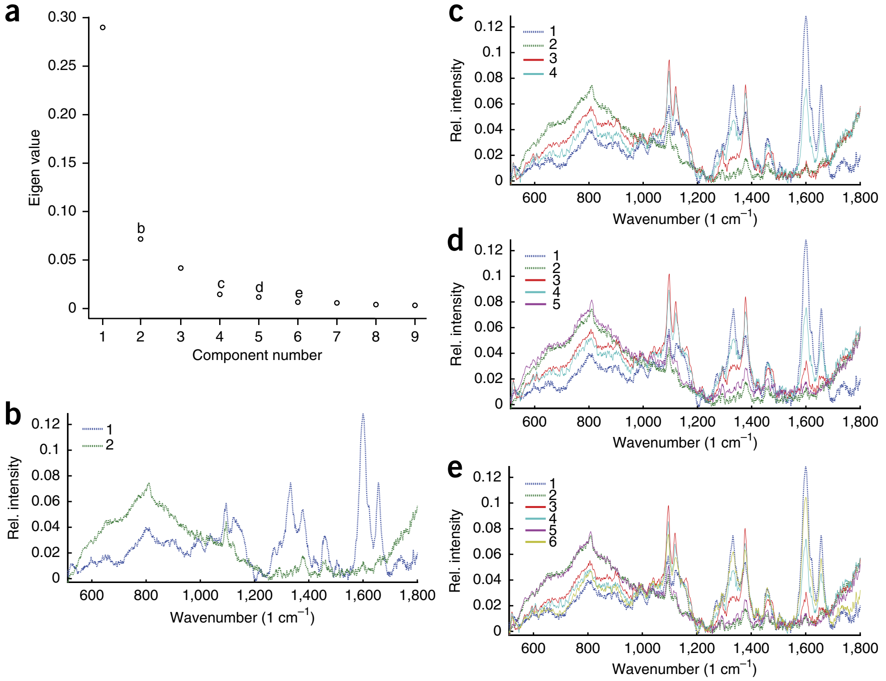

Figure 7 | Selecting the number of components for MCR-ALS, based on singular value decomposition (SVD). (**a**) The 'Singular Value Decomposition' plot of the main interface, which shows the singular values in a typical Raman hyperspectral image of wood fibers in the cross-section of a hybrid aspen stem. The largest drop is seen between the first and second singular values (marked b). Singular values level off somewhat at the fourth component (mark c), with a minor drop between the fifth (mark d) and sixth (mark e) components. Thereafter, singular values change only marginally. On the basis of these values, two, four, five or six components can be selected for MCR-ALS (marks b, c, d and e, respectively). The corresponding 'Pure Spectra Estimate (Initial Values)' plots are shown in **b**–**e**, and they are used to help deciding the number of components for MCR-ALS. (**b**) Two-component model: the spectrum for component 1 (blue) shows clear bands, whereas the spectrum for component 2 (green) only contains contribution from fluorescence baseline and noise. This indicates that using only two components would likely resolve only cell wall and lumen differences (i.e., showing the presence or absence of sample). (**c**) Four-component model: all four spectral profiles are unique, and all but one (for component 2, green) also contain clear bands. This indicates that all components are unique, and describe different chemical components in the cell wall (and the lumen, component 2, green). (**d**) Five-component model: two of the spectra are practically identical (components 2 and 5, green and purple, respectively) and only represent fluorescence baseline and noise. This indicates that there is likely no need for the fifth component. (**e**) Six-component model: although the spectrum of component 6 (yellow) contains real bands, it is essentially identical to the spectrum of component 1 (blue), except for generally lower intensity. Thus, it is not unique. As seen in the five-component model already, the spectrum of component 5 is identical to that of component 2, and it only describes fluorescence baseline and noise. This indicates that there is probably no need for the sixth component either. On the basis of the above evaluations, the best option is to use four components in the subsequent MCR-ALS analysis.

the biological, chemical or spectroscopic evaluation finds it necessary and meaningful (see the 'Component profiles (biological, chemical and spectroscopic evaluation of concentration maps and spectral signatures)' section below).

Lengthening the iteration process either by increasing the number of iterations or by decreasing the convergence criterion is only helpful if very few iterations were initially used. In other GUIs that use the MCR-ALS algorithm with more constraint options[40], a high lack of fit can also be associated with constraints that are improper or too strict. However, this is not the case in the simplified GUI presented in this work, as the non-negativity constraint should already be met after pre-processing.

Component profiles (biological, chemical and spectroscopic evaluation of concentration maps and spectral signatures). After performing MCR-ALS, the concentrations of the resolved pure components should be mapped. These distribution (concentration or component) maps should be evaluated to find matches to features observable in the visible image, global intensity plots or with any a priori biological knowledge (biological evaluation). In the best case, these component maps can already indicate zones of distinct chemical composition. In this respect, they provide an analytical endpoint, agreeing with or providing complementary information to segmentation maps. However, random maps or maps that do not reflect the biological features of the sample can be artifacts or products of inadequate pre-processing or of data quality or quantity (**Fig. 6**).

Together with the component distribution maps, the spectra of the corresponding resolved components should also be inspected (chemical and spectroscopic evaluation). Spectral shapes should be generally meaningful and never completely different from the features of the raw spectra. Ideally, the spectra should contain characteristic bands of certain compounds in image areas covered by sample, and only background or signal unrelated to the biological material in areas that are free from the sample (e.g., glass or embedding media in the cell lumen or at tissue boundaries, **Fig. 2**). Although artifacts in theory could also be resolved as components, these should be eliminated by the pre-processing step (see the strong fluorescence signal in the beginning of the recording, which diminishes during the scan in **Fig. 6**). Such artifacts are easily detected in both the component maps and the corresponding spectra, clearly indicating improper pre-processing.

Reference spectra matching (chemical evaluation). Optionally, the spectra of the resolved components can be compared with reference spectra for identification (chemical evaluation). Although the pure components identified by MCR-ALS are often pure compounds (or very good approximations of them), some can still have a mixed character (i.e., including signal from more than one chemical compound). This can be because of the absence of more powerful (e.g., local rank[46,47]) constraints in the data analysis pipeline, or simply because certain conditions of compound overlap are not fulfilled by the sample. The reference spectra to be matched to the resolved pure components should be recorded separately by using well-defined chemicals (e.g., different wood biopolymers extracted by wet chemical methods, relevant model compounds or mixtures with known compositions), or extracted from well-defined regions of other images for investigating similarities.

One of the most common methods for spectral matching is based on Euclidean distances (dot products), because it is computationally undemanding yet robust[48–50]. In a simplified view, comparison is based on determining the distance (difference) between each data point of the component and the reference spectra. The closer they are to each other, the higher the match is. Such point-by-point matching requires the reference spectra to have the exact same number of data points as the sample it is matched to. In spectroscopic terminology, it translates to covering the same spectral region with the same spectral resolution. In addition, spectra should be as similar as possible in terms of minimum and maximum intensity, so data points are not distant simply because of offset or intensity differences. For this reason, resolved pure component spectra and loaded reference spectra are automatically offset-corrected and area-normalized by the script presented in this protocol. To further facilitate matching, the loaded reference spectra can be pretreated in the same way as the raw input data, using the same parameters for baseline correction and smoothing.

It must be pointed out that reference matching is entirely optional and that the results must always be carefully evaluated. On one hand, Euclidean distances can produce relatively high matches even in cases in which spectra are obviously different (false-positive match). Consider the following hypothetical scenario: both the component and the reference spectra have a single characteristic band each, albeit in different positions, or they have only small uncharacteristic bands or large regions with only baseline. In this case, the two spectra will be rather closely matching at every data point except for the two small zones in which their respective characteristic bands are located. Such false-positive matches are easily identified by visual inspection (**Supplementary Fig. 4**). On the other hand, substantial portions of the spectra can be different, resulting in low Euclidean matches. However, this may simply be due to the fact that the pure components are not pure compounds but unresolved mixtures. In this case, a low match indicates a low content of the reference compound in the mixture, rather than a poor match (**Supplementary Fig. 4**). Although in a sense it provides a false-negative match, it is harder to detect than the false-positive match in the hypothetical example above. Imperfect matches can also originate from reference spectra that do not mimic exactly the pure spectrum of a compound in the particular biological sample analyzed. For further notes on reference matching, consult the 'Additional comments and limitations' section below.

Segmentation (biological and chemical evaluation). One of the challenges in hyperspectral imaging is the identification of equivalent pixels between independent images so that appropriate comparisons can be made in statistical tests. Component maps often show gradients of their respective components, and thus they cannot always be used to find clear boundaries between chemically different areas. In addition, zones based on multiple component profiles cannot easily be identified between independent images. Although manual categorization on a pixel-by-pixel basis is time-consuming and subjective, image segmentation can be performed automatically and objectively by *k*-means clustering of the MCR-ALS concentration profiles[34]. The key parameter for *k*-means clustering is the number of clusters, which can be determined manually or iteratively using silhouette values. The silhouette value of each data point (pixel spectrum) in the image describes how similar that pixel is to other pixels in its own cluster compared with pixels belonging to other clusters. The similarity is determined on the basis of the average difference from pixels in the same cluster versus the minimum average difference to pixels in a different cluster[28]. There are numerous measures of such differences[28], including Euclidean distances. Nevertheless, the number of clusters determined by silhouette values can always be manually overruled in both directions—i.e., fewer or more clusters can be selected for *k*-means clustering on the basis of a priori knowledge of the sample or biologically justified expectations. In general, it is recommended to test different numbers of clusters and compare the results. The resulting segmentation maps, for instance, should be compared with the white light image to evaluate whether they match anatomical features or sample areas of interest to the biologist (biological evaluation). In addition, the contribution of each pure component to a particular cluster (image segment) should be investigated by means of the corresponding centroid plot (chemical evaluation). This strategy is useful to clearly identify chemically different regions with distinct boundaries in a spatial and compositional manner, and it can be more powerful than heat maps of single-band intensities or pure component maps. Spectra from equivalent regions of multiple images can thus be compared directly, or by using multivariate methods, such as orthogonal projections to latent structures–discriminant analysis (OPLS-DA)[7,29].

Additional comments and limitations

Our GUI is based on modules that were primarily designed to handle a broad set of analytical challenges (multiple methods from chromatography to spectroscopy, multiple series of spectra, multiple images)[26,40]. To enhance the ease of use and to streamline the analysis for a specific task only (namely vibrational microspectroscopic image analysis of biological materials), certain parameters have been fixed, preventing end-user modifications and excluding certain analytical options altogether. Here we outline the key limitations of this protocol. Consultation with spectroscopy and chemometrics experts using **Table 1** may be necessary to determine whether these limitations make the protocol unsuitable for a specific task.

In particular, non-negativity constraints in the present script are fixed for both spectra and concentrations in MCR-ALS. The selection of the number of components is based on SVD or manual input only, and starting conditions are locked to pure spectral estimates using a method based on SIMPLISMA[43]. The spectra in each pixel must have equal spectral range and resolution. Moreover, the presented script does not allow simultaneous multitechnique or multi-image processing. However, it does allow for spectral pre-processing, including AsLS baseline correction, optional smoothing and area normalization. AsLS baseline correction is limited to spectra with non-negative peaks, and thus it cannot handle derivatives. Smoothing is only available by the S-G method.

In addition to the practical constraints of MCR-ALS listed in **Table 1**, some aspects should be taken into account. MCR-ALS is often applied in a dynamic way to test different sets of starting parameters until satisfactory final results are obtained. It is important to note that MCR-ALS is not a nested method; i.e., increasing the model size does not imply adding new components to the existing ones. Instead, it means that the full system is re-defined with a new set of components. Therefore, it is recommended to test different number of components (**Fig. 7**) on the same image to ensure optimum unmixing. Another way to influence the optimization is by using local rank constraints[40,46,47]. Although these constraints are highly valuable in cases of extreme compound overlap (to better define the presence and absence of compounds in pixels), they are very difficult to automate. In biological materials, in which the different elements are reasonably well compartmented (tissues, cells, subcellular zones), MCR-ALS using only non-negativity constraints often provide very satisfactory results without the need of local-rank constrains.

As MCR-ALS is not used with all its capabilities in the present script, unmixing may not be perfect in some cases. Thus, the resulting pure components may be good approximations of pure compounds, but still mixtures (i.e., containing spectral signatures from more than one chemical compound). Reference matching is helpful in this respect, as it can show the dominating compounds in the pure component spectra, but it has its own limitations. Most importantly, the separately recorded reference compounds may not be chemically identical to the native *in situ* compounds, which are integrated into biological systems and linked to other biomolecules (differences in, e.g., cellulose crystallinity, protein structure and folding, and polymer orientation). This is often reflected in their spectra and results in less-than-perfect matches. It is important to note that reference spectra must be recorded with similar settings to the samples, having identical spectral region and resolution. If needed, reference spectra should be pre-processed in the same way as the image (i.e., baseline correction and smoothing).

When these limitations render this protocol unsuitable for a particular data set, we refer to a more complete MCR-ALS GUI with additional MATLAB scripts[40], focusing on the resolution data analysis step. It offers more versatility and can be used with a wider range of constraints to a broad variety of data analysis problems (including simultaneous multitechnique and multi-image processing), but it also requires a better understanding of chemometrics.

Finally, as a practical limitation, the input data for our GUI must be either an ASCII file (e.g., a common tab- or space-delimited .txt file) or a MathWorks MATLAB .mat file. For requirements regarding files, folders and formatting, please refer to the PROCEDURE section.

Experimental details

Biological material. The materials that we chose to illustrate the different steps of our method are cryo-sections from xylem (wood) of deep-frozen hybrid aspen (*Populus tremula* × *Populus tremuloides*) stems. Raman image data were recorded by available standard procedures[14]. Small maps (<200 pixels) were used to demonstrate the suitability of our method for screening large sample numbers where recording times of individual images need to be kept short (and therefore pixel numbers kept low). We illustrate this by showing how our pipeline can be used to distinguish the chemically distinct cell wall layers in such images. However, our protocol is versatile and well suited to different kinds of vibrational (FTIR or Raman) microspectroscopic images of diverse sample types and image sizes (pixel numbers). To demonstrate this versatility, we provide a further example using an FTIR microspectroscopic image of a biomedical mammalian tissue (the pancreas of 6-week-old C57BL/6 mouse; **Fig. 2**).

Input data. The following input data files were used in the protocol (and are available as **Supplementary Data**, including the corresponding white light images in .jpg format):

- BadExample_Aspen_Raman.txt is a Raman microspectroscopic image with substantial fluorescence problems and limited number of pixels. It was recorded over normal wood in hybrid poplar (see 'Biological material' above), using a Renishaw inVia microscope, 100× magnification, 514-nm Ar$^+$ laser excitation, 1-μm laser spot and step size and 1-s exposure time, in static mode. Please note that this bad example was deliberately chosen to illustrate problems, and that it does not indicate a limited suitability to analyze normal wood samples in general.
- GoodExample_Aspen_Raman.txt was recorded with the exact same settings as BadExample_Aspen_Raman.txt, over the tension wood area of hybrid poplar.
- The corresponding white light images of the Aspen examples (**Supplementary Data**) contain overlaid single-band intensity heat maps, marking the area scanned by Raman microspectroscopy. These heat maps were created using the built-in functions of Renishaw's WiRE software, calculating the band total band area in the 1,560–1,650 cm^{-1} region, corresponding to aromatic −C=C− vibrations (lignin).

For **Figure 6**, spectra were pre-processed using AsLS baseline correction ($\lambda = 30{,}000$, $P = 0.001$) and S-G smoothing (order = 1, frame = 3), before performing MCR-ALS. MCR-ALS modeling was performed using five components, and default parameters (50 iterations, 0.1 convergence limit). Segmentation was performed using four clusters for both examples in order to compare results, although silhouette clustering for the 'bad example' image (**Supplementary Data**) returned only two clusters.

The reference compound spectra (cellulose.txt, lignin.txt and Dglucuronicacid.txt; **Supplementary Data**) were also recorded with the exact same settings as BadExample_Aspen_Raman.txt, using the same chemicals as described in Gorzsás *et al.*[7].

064 × 064_MousePancreas_FTIR.mat (**Supplementary Data**) is an FTIR microspectroscopic image, which is recorded over the wax-embedded pancreas section of a 6-week-old C57BL/6 mouse, using a Bruker Tensor 27 FTIR spectrometer with an attached Hyperion 3000 microscopy accessory and a 64 × 64 liquid nitrogen–cooled focal plane array (FPA) detector. The recorded area covers ~175 × 175 μm in 64 × 64 pixels, resulting in a physical pixel dimension of ~2.7 × 2.7 μm. Lateral resolution, however, is diffraction-limited, and it ranges from ~10 μm to 20 μm in the 1,600–1,000 cm^{-1} spectral range, depending on the wavelength of the light. 32 interferograms were coadded for improved signal-to-noise ratio. Spectral resolution of 4 cm^{-1} was used and a zero filling factor of 2 was applied.

For **Figure 2**, spectra were pre-processed using AsLS baseline correction ($\lambda = 100$, $P = 0.001$) and S-G smoothing (order = 1, frame = 3), before performing MCR-ALS. MCR-ALS modeling was performed using five components and default parameters (50 iterations, 0.1 convergence limit). Segmentation was performed using three clusters, as the image is expected to contain empty areas (covered by the embedding medium), cells of the exocrine tissue and cells of the islets of Langerhans.

MATERIALS

EQUIPMENT

Hardware

- Standard-equipped PC or Mac with minimum system requirements to run the software (see below) and enough free disk space for saving the results. For optimal viewing of the interface, a screen resolution of 1,920 × 1,200 is recommended. Low-resolution screens can result in cropping of text, in which case please refer to **Supplementary Figure 1** for the full text. MCR-ALS results as text are also displayed in the MATLAB Command window, in case they are cropped in low-resolution screens (**Table 2**)
- For the TIMING sections in the protocol, the following architectures were used: Apple MacBook Pro 15-inch Retina, 2.4-GHz Intel Core i7 CPU, 8 GB 1600 MHz DDR3 RAM, Intel HD Graphics 4,000 1,024 MB, 256 GB SDD, OS X 10.9.1; Fujitsu Siemens Celsius desktop, 2.67-GHz Intel Xeon CPU, 8 GB RAM, NVidia GeForce 6600 256 MB, 1TB HDD, Windows 7 Professional (64-bit); Apple MacBook Air 13-inch, 2.13-GHz Intel Core 2 Duo, 2 GB 1,067-MHz DDR3, NVIDIA GeForce Graphics 9400M 256 MB, OS X 10.9.1;Apple iMac 22-inch, 2.7-GHz Core i5 20GB 1333 MHz DDR3, AMD Radeon HD 6770M 512 MB, OS X 10.9

Software

- MathWorks MATLAB version R2012a or newer is recommended. The GUI has been successfully tested on MATLAB version R2009b with no problems. However, older versions are not supported
- MCR_ALS_PlantImaging_v1.m and .fig files (available for download in a single .zip archive, free of charge at http://www.kbc.umu.se/vibrationaldownload.html, by selecting the 'MCRALS imaging' download link)
- For the TIMING sections in the protocol, the following MATLAB versions have been used: Version 8.0.0.783 (R2012b), Mac (64 bit); version 8.2.0.701 (2013b), Win (64 bit); version 8.2.0.701 (2013b), Mac (64 bit)

EQUIPMENT SETUP

Input data: general considerations For the GUI to work, the input files must be either tab- or space-delimited ASCII files (e.g., .txt files), or MATLAB .mat files. Other file types or files with a data structure that is different from that outlined below need to be re-formatted to match the input file requirements of the GUI. For better organization, keep the files in their dedicated folder without unnecessary nested subfolders to facilitate navigation. The image files used as examples in the present protocol are provided as **Supplementary Data**, together with three reference spectra files. These can be used for testing the protocol and as formatting guides for own data.
! CAUTION Folder and filenames must contain standard alphanumeric characters only (i.e., unaccented Latin letters, numbers and underscore). No special characters are allowed. Do not use capitals; only use small letters for file extensions (i.e., .jpg and not .JPG, .txt and not .TXT, .mat and not .MAT and so on), as MATLAB is case-sensitive.

Format for ASCII files

ASCII files must have the following format:

```
1 1 1800 4528
1 1 1799 4321
1 1 1798 4413
 . .  ....  ....
1 2 1800 4619
1 2 1799 4721
1 2 1798 4812
. .  ....  ....
1 3 1800 4311
1 3 1799 4289
1 3 1798 4156
. .  ....  ....
. .  ....  ....
2 1 1800 4555
2 1 1799 4369
2 1 1798 4611
. .  ....  ....
```

In this case, the first column contains the *y* coordinate of the pixel, the second column contains the *x* coordinate of the pixel, the third column contains the wavenumbers and the fourth column contains the spectral intensities. This is the default format when Raman image maps are recorded by Renishaw's WiRE software and exported as .txt files.

Format for MATLAB files

MATLAB .mat files must have the following structure:

```
4000 0.0473 0.0498 0.0953 ... 0.0622
3998 0.0423 0.0123 0.1022 ... 0.0817
3996 0.0494 0.0331 0.1169 ... 0.0724
.... ...... ...... ...... ... ......
400 0.0512 0.0678 0.0744 ... 0.0688
```

In this case, the first column contains the wavenumbers; thereafter, each column contains the intensities of one pixel. The first pixel is the one located in the first row of the first column of the image, the second pixel is the first row of the second column and so on. This is the default format with Bruker's OPUS 7 export into MATLAB. **! CAUTION** The .mat file does not contain information regarding *x* and *y* coordinates; it only contains the total number of pixels (i.e., the .mat file of an image containing 4 × 20 pixels has the same dimensions as the .mat file of an image containing 8 × 10 pixels). Thus, the number of *x* and *y* pixels needs to be supplied separately when .mat files are used as input. Alternatively, the filename can contain this information in the following format: 'AAA × BBB_examplefilename.mat'. In this case, AAA and BBB denote the number of pixels in the *x* and *y* dimensions, respectively. For instance, 064 × 032_examplefilename.mat indicates an image with 64 × 32 pixels. **! CAUTION** As .mat input files are processed faster than .txt files, they are recommended for larger data sets (>200 pixels).

Reference spectra (optional) Input files for reference spectra must be tab- or space-separated ASCII files (e.g., .txt files), with the first column containing wavenumbers and the second column containing the corresponding intensities. No other formats are accepted.

PROCEDURE

▲ **CRITICAL** Many of the steps are prefaced by the term 'optional'. This means that technically future steps are not dependent on these having been performed. However, they can provide additional information, or they can alter the outcome of the analysis. These optional steps are possible actions available to the user in the GUI to experiment and find the best settings on a case-by-case basis.

Starting steps ● TIMING ~30 s–5 min

1| Start MATLAB.

2| Navigate to the folder containing the 'MCR_ALS_PlantImaging_v1.m' script, using the 'Current Folder' panel in MATLAB.

3| Double-click on the MCR_ALS_PlantImaging_v1.m file in the 'Current Folder' panel to load the script to the main MATLAB interface.

4| Run the script. This step can be performed using option A or B, depending on personal preference.

(A) 'Run' button

(i) In the MATLAB interface, under the EDITOR tab, click on the 'Run' button.

(B) Right-click

(i) Right-click on the file MCR_ALS_PlantImaging_v1.m and select 'run' in the opening context-sensitive menu.

▲ **CRITICAL STEP** If the MCR_ALS_PlantImaging_v1.m file is not in the default MATLAB folder but it is automatically loaded at startup (i.e., it was left in the Editor window the last time MATLAB was shut down), MATLAB will ask whether to change to that folder or add that folder to the path. In this case, select changing to that folder.

? TROUBLESHOOTING

5| In the 'Open File' dialog box, navigate to the hyperspectral image data file to be processed, select it and click 'Open'.

▲ **CRITICAL STEP** If a .mat file was selected with no pixel dimension defined in the filename, an additional dialog box will open, where the correct *x* and *y* pixel numbers need to be supplied manually.

▲ **CRITICAL STEP** Data need to comply with the formatting rules outlined in the Input data section of this PROCEDURE.

? TROUBLESHOOTING

6| The data are loaded and the relevant parts of the GUI are automatically populated. Before going forward with the remaining procedure, have a look at the data displayed on this screen (**Box 1**). At initial startup, displaying the 'Corrected Spectra' plot may be slow, requiring up to several minutes to load large data files on slow computers.

! CAUTION Cropping of text in the GUI can occur on low-resolution computer screens.

? TROUBLESHOOTING

7| (Optional) Save any or all of the three plots 'Original Spectra', 'Selected Spectrum' and 'Corrected Spectra' by clicking on their respective 'Save Plot' buttons, located on the top right of each plot. Suggestions for file type (.pdf by default), filename and folder are prefilled in the opening save dialog boxes, but they can be changed.

? TROUBLESHOOTING

8| (Optional) Save the total intensity plot by clicking on the 'Save Intensity Map' button, located to the left of the plot. Suggestions for file type (.pdf by default), filename and folder are prefilled in the opening dialog box, but they can be changed.
? TROUBLESHOOTING

9| (Optional) Manually load a white light image by clicking on the 'Load White Light Image' button in the 'Visualization' frame. Use the opening dialog box to navigate to the correct folder and select the file containing the white light image.
? TROUBLESHOOTING

Pre-processing steps ● TIMING ~1–20 min
▲ CRITICAL If the data have already been correctly pre-processed previously, and they are only reloaded for MCR-ALS analysis, pass the input data directly to MCR without performing pre-processing (i.e., bypassing baseline correction, smoothing and area normalization, **Fig. 1**) by clicking on the 'Untreated for MCR' button. In this case, ignore Steps 10–16 and proceed from Step 17.

10| Perform baseline correction using AsLS fitting[26]. Adjust the baseline using either the sliders or the direct input text boxes for λ and P values in the 'Baseline' frame of the interface.
▲ CRITICAL STEP The aim is to completely remove the baseline while retaining all spectral information (**Fig. 3a**). Generally, the best strategy is to keep the P value to the minimum, as there should be no negative peaks, and they only vary the λ value. Testing λ values is recommended in the range of 10^2–10^9, varying one order of magnitude among successive testing steps to clearly see the effect on the fitted baseline shape[26].
▲ CRITICAL STEP Too-high λ values result in underfitting (i.e., not all of the baseline is removed, **Fig. 3c**), whereas too-low λ values result in overfitting (i.e., spectral bands are removed as well, and not only baseline, **Fig. 3b**).
? TROUBLESHOOTING

11| Confirm that the baseline correction performs equally well on different parts (pixels) of the image by selecting different spectra using either the drop-down menu or the direct input text box above the 'Selected Spectrum' plot.
? TROUBLESHOOTING

12| (Optional) Tick the checkbox 'S-G filtering' in the 'Smoothing' frame of the interface to perform smoothing of the spectra using S-G filtering[27]. The 'Selected Spectrum' plot automatically updates and shows the smoothed spectra in green.
▲ CRITICAL STEP The polynomial order (in the box labeled 'Order') must be lower than the frame size (in the box labeled 'Frame'), which must be an odd number. The larger the difference between Order and Frame, the more smoothing will be applied. If order = frame – 1, no smoothing is performed[28].
▲ CRITICAL STEP The use of too much smoothing results in the loss of spectral resolution, distortions and merging of bands (**Fig. 4**).
? TROUBLESHOOTING

13| (Optional) Tick the checkbox labeled 'Area Norm' to perform area (total intensity) normalization over the entire spectral region (see the INTRODUCTION section and **Supplementary Fig. 2**).

14| Click on the 'Update Plot' button above the 'Corrected Spectra' plot to check the combined outcome of all spectral pre-processing steps (Steps 10–13) on the spectra. If it is suboptimal, return to Step 10.
▲ CRITICAL STEP Spectra need to be free from physical and optical artifacts as much as possible before proceeding to the following steps. Otherwise, these artifacts can be resolved as pure components by MCR-ALS and distort the results (**Fig. 6**).

15| (Optional) Save the pre-processed data in the same format as the original input file (ASCII or .mat) by clicking on the 'Save Corrected' button in the bottom left of the interface. Saving in ASCII .txt format can be slow for large data sets (large spectral range and many pixels). In those cases, .mat formats are recommended as input.

16| Click on the 'Pre-Treated for MCR' button to conclude all pre-processing steps, and send the data for MCR-ALS analysis. Clicking this button automatically updates the Total Intensity Map in the 'Visualization' frame on the top right of the interface (**Box 1**, step 5) and populates the SVD plot in top middle of the interface (**Supplementary Fig. 1**) in the 'Multivariate Curve Resolution' frame. The updated Total Intensity Map considers baseline changes and smoothing, but it does NOT consider area normalization, as that would result in the loss of features owing to all intensities being equal after normalization.

Data analysis steps ● TIMING ~2–30 min
17| Decide how many components you want the program to consider. The choice needs to be based on the results of SVD (top plot in the 'Multivariate Curve Resolution' frame in the central part of the interface, **Fig. 7**), or on a priori knowledge and expectations (see the INTRODUCTION and **Fig. 7**).

Box 1 | Initial view of the data in the GUI

1. The name of the loaded data file is displayed in red in the bottom right part of the interface below the 'Segmentation Map (Clusters)' and 'Centroid Profiles' plots.
2. The top left plot in the interface labeled 'Original Spectra' (in the 'Pre-processing' frame) shows the loaded spectra in a rainbow set of colors and without modification.
! CAUTION To speed up the displaying step on slower computers, this plot contains a maximum of 100 spectra. If the original data set contains more spectra, 100 of them are randomly selected for display.
3. The 'Selected Spectrum' plot displays in red the currently selected spectrum in the data set (by default the first one or the first randomly selected one if there are more than 100 spectra in total), the AsLS calculated baseline in blue (using the default settings, Step 10) and the resulting corrected spectrum in black. The drop-down menu above the plot shows the number of the currently selected spectrum.
? TROUBLESHOOTING
4. The 'Corrected Spectra' plot in the bottom left of the interface shows the result of the pre-processing with the set parameters. It displays the same set of spectra as shown in the 'Original Spectra' plot, using the exact same color for each spectrum.
5. The 'Total Intensity Map' in the top right of the interface, under the 'Visualization' frame, shows a total intensity map created by determining the area under all bands in the entire spectral region. The colors range from dark blue (lowest intensity) to dark red (highest intensity).
This plot uses the raw data input with no baseline correction at startup. The plot automatically updates once the pre-processed data are submitted for MCR-ALS analysis (Step 17).
6. A white light image is automatically loaded as long as it is located in the same folder and has the exact same filename as the data file loaded, with .jpg as extension. It is shown in the 'White Light Image' plot in the top far right of the interface, in the 'Visualization' frame.
Certain button texts in the GUI are color-coded to differ from the default black. Red text means that the button starts an important stage of the analysis, whereas blue text means that the button saves the results of an important stage of the analysis.
? TROUBLESHOOTING

18| Select the number of components by either of the following options. The first is to use the drop-down menu of Eigen values, listed in ascending order of component number (lower Eigen value for higher component number). The text box below automatically updates to show the selected number of components in red. The second option is to type the number of components into the text box with the same name and hit enter. The drop-down menu above automatically updates to show the corresponding singular value.

After selecting the number of components, the 'Singular value decomposition' plot automatically zooms in to show the selected number of components +5. (For example, selecting four components, the plot will show the first nine, i.e., 4 + 5, components, **Fig. 7a**). In addition, selecting the number of components automatically populates the 'Pure Spectra Estimation (Initial Values)' plot with initial pure spectral estimates, which is calculated using a SIMPLISMA-based method (see INTRODUCTION). The 'Purest Pixels' list automatically updates to show the pixel number of the pure spectral estimates (Step 19).

▲ CRITICAL STEP Selecting the correct number of components ultimately determines the outcome of the analysis. Different numbers of components should be tested and the results should be evaluated to see which gives the best result in terms of chemistry and biology (see INTRODUCTION).

! CAUTION Adding a new component is not an additive process. In other words, it does not leave the original components intact, but it recalculates all components (see INTRODUCTION).

? TROUBLESHOOTING

19| Inspect the 'Pure Spectra Estimation (Initial Values)' plot to see whether spectral profiles are reasonable or not (e.g., whether they contain only noise or are very similar to each other (can indicate too many components selected), whether they show every band in the spectra with equal weight (can indicate too few components selected), or whether they contain spectral artifacts (can indicate improper pre-processing)).

▲ CRITICAL STEP It is important to see whether the pure spectral estimates are meaningful or not, as this can help in selecting the correct number of components (**Fig. 7** and INTRODUCTION). If the addition of a new component (Step 18) does not result in a significantly different new spectral estimate, it is likely that the new component is not required and will not be well resolved (see 'BACKGROUND' section).

▲ CRITICAL STEP The colors used for each component in the 'Pure Spectra Estimation (Initial Values)' plot are kept constant throughout the interface. Thus, the legend of this plot can always be consulted to determine which color represents which component in subsequent plots. The only exceptions are plots containing reference spectra (Steps 36 and 38).

? TROUBLESHOOTING

20| (Optional) Change the noise allowed (in percentage) for the calculation of initial estimates in the text box labeled 'Noise' and hit enter. Although the default value of 10% is generally safe, different values can be tested and their effect inspected in the 'Pure Spectra Estimation (Initial Values)' plot.
? TROUBLESHOOTING

21| (Optional) Mark the location of the purest pixels on the Total Intensity Map by ticking the checkbox labeled 'Mark Purest' in the Visualization frame of the interface, to the left of the Total Intensity Map.
? TROUBLESHOOTING

22| Use the location of these pixels in the image together with the 'Pure Spectra Estimates (Initial Values)' plot (Step 19 and **Fig. 7**) to evaluate whether the correct number of components has been selected or not.
! CAUTION Once this step has been reached, the locations of the purest pixels are not updated automatically after returning to Step 18 to select a different number of components. To force the locations of the purest pixels to update, untick and tick the 'Mark Purest' checkbox. However, this only works if a higher number of components is selected compared with that selected previously in Step 18. If a lower number of components is selected, the 'Mark Purest' checkbox is unable to properly display the fewer components, and the interface needs to be restarted by performing Steps 45–48 and returning to Step 4.

23| (Optional) Click on the 'Show Eigen Vectors' button to the right of the 'Singular value decomposition' plot to show the Eigenvectors of the SVD in a separate window. Eigenvector profiles showing a noisy random pattern refer to nonrelevant contributions. This option is especially valuable for people with more thorough background in chemometrics or when more insight is required, as distinction between noisy pattern and minor spectral features is not always obvious. As such, it can be safely omitted in general practice.

24| (Optional) Before starting the MCR analysis, you can have a quick look at the description of the data set by bilinear models based on components. Click on the 'Data Preview' button to perform a PCA analysis; this opens a separate window that shows four different plots. The first and second plots refer to an approximate description of the data set with a bilinear model of meaningful contributions obtained with the 'Initial Pure Spectra Estimates' (first plot), and the related concentration profiles that are estimated using the original data set and the spectral estimates by least squares calculation without any constraints (second plot). The third and fourth plots show the bilinear model obtained by PCA, using the same number of components as set for the MCR analysis. In addition, the main Command Window of MATLAB displays the PCA analysis parameters (number of components, lack of fit and CPU time). These plots provide a first insight into the behavior of the data set that will be improved with MCR analysis. Although informative, this step is not strictly necessary to perform the MCR analysis and is therefore optional.

25| Set the parameters for MCR-ALS. As most constrains are already preset (**Table 1**), the two main parameters to be adjusted at this stage are the number of iterations and the convergence limit. The convergence limit determines the σ change between consecutive iterations (improvement of fit), below which the solution is considered to be optimal and is therefore not refined further. The number of iterations sets the maximum number of iterations performed, unless convergence is achieved before. In practice, it is difficult to know the optimum values for these parameters beforehand, and therefore the default values can be used at the start.
▲ CRITICAL STEP High convergence limits or low numbers of iterations will result in quicker (and perhaps suboptimal) analysis, which can be preferred for a quick overview.
▲ CRITICAL STEP The live updating MCR-Optimized Spectra and Concentration plots (Step 26) are the most visual indicators to determine whether the convergence limit and the number of iterations are set properly or not. If either of these plots changes significantly even at the last iteration, lower the convergence limits and/or increase the number of iterations. Similarly, if the 'MCR Results' displays anything other than 'CONVERGENCE ACHIEVED' (Step 26), adjusting the convergence limit and/or the number of iterations may be needed. If a warning of divergence appears, the way in which the MCR analysis has been set (number of components, initial estimates and so on) and the initial submitted data set (pre-processing options) should be reconsidered (return to Step 18 or Step 10, respectively).

26| Perform MCR-ALS analysis by clicking on the 'Perform MCR' button. The analysis is performed with the preset constrains (see **Table 1** and the INTRODUCTION) and parameters set in Step 25. The MATLAB Command Window and the 'MCR Results' frame in the center of the interface both display the results at every step of iteration. The first line of the display shows the current number of iterations ('Iterations:'). Below this is the status of the analysis, which can be 'FIT IS IMPROVING' or 'FIT IS NOT IMPROVING', resulting in 'CONVERGENCE ACHIEVED' (when the change in σ values is below the convergence limit set in step 25), 'FIT NOT IMPROVED 20 TIMES. STOP' (when the fit has not improved for 20 consecutive iterations, indicating divergence) or 'MAX NR OF ITERATIONS' (when the iteration count reaches the maximum allowed in Step 25, without reaching convergence or stopping for divergence). The lines following the status display show the percentage change in σ values, the fitting errors expressed as percentage lack of fit for PCA and the experimental variation, and finally the percentage of variation explained. In addition to these parameters, the MATLAB Command Window also lists the sum of squares in PCA reproduction, the sigma with respect to the experimental data and it finally gives a summary line at the end of the iteration. Of these parameters, the percentage σ change can be useful for adjusting the convergence limit in Step 25 for subsequent analysis, whereas the lack of fit and variance explained are indicative of how well the current model describes the data (see the 'BACKGROUND' section).

▲ CRITICAL STEP A model with low fit and low explained variance cannot be considered a good representation of the data. If these values cannot be improved by changing key pre-processing parameters (return to Step 10) or MCR-ALS parameters (such as the number of components (return to Step 18), noise in initial estimates (return to Step 20) or maximum number of iterations and convergence limits (return to Step 25; **Fig. 1**)), consultation with chemometrics and spectroscopy experts is necessary.

▲ CRITICAL STEP During the MCR-ALS analysis, the plots 'MCR Optimized Spectra' and 'MCR Optimized Concentration' automatically update with each iteration. These plots (especially the spectral output) are important for evaluating the results and to see whether there is any major change during the iterations of MCR-ALS. If changes are still significant at the last iteration, the convergence limit and/or the maximum number of iterations need to be adjusted (return to Step 25).

? TROUBLESHOOTING

27| (Optional) Save the 'MCR Optimized Spectra' and 'MCR Optimized Concentration' plots using their respective 'Save Plot' buttons, located above each plot. Suggestions for file type (.pdf by default), filename and folder are prefilled in the opening save dialog boxes, but they can be changed.

? TROUBLESHOOTING

28| (Optional) Save the 'MCR Optimized Spectra' and 'MCR Optimized Concentrations' matrices using their respective 'Save Matrix' buttons, located above each plot. Suggestions for filename and folder are prefilled in the opening save dialog boxes, but they can be changed. The resulting .mat files contain the variables *SOpt* and *COpt*, respectively. The variable SOpt stores the optimized spectral profiles of the pure resolved components in $N \times W$ format, where N is the number of components and W is the number of wavenumbers (i.e., each row represents one spectrum, see S^T in **Fig. 5**). The variable *COpt* stores the optimized concentration profiles of the pure resolved components in $M \times N$ format, where M is the number of total pixels and N is the number of components (i.e., each column contains the concentration of a single pure component in all pixels; see C in **Fig. 5**).

Visualization steps (evaluation) ● TIMING ~10 s–1 min

29| (Optional) If no white light image is loaded automatically at startup (**Box 1**, step 6), or if a new one needs to be loaded instead of the default one, click on the 'Load White Light Image' button in the 'Visualization' frame in the upper right of the main interface. A standard load dialog box opens, which allows for navigation and single image selection, listing only .jpg files by default. Multiple images cannot be selected and loaded.

? TROUBLESHOOTING

30| (Optional) Tick or untick the 'Mark Purest' checkbox to show or hide the purest pixels in both the Total Intensity Map and in the Component Maps (Steps 21 and 31). This can be useful to see which regions in the image are associated with each pure component.

! CAUTION The marks only show locations, but they do not display which pure component is associated with a certain pixel. They do not show pixel numbers either.

? TROUBLESHOOTING

31| Click on the 'Show Component Maps' button in the 'Visualization' frame in the upper right section of the interface. This opens a new window, which contains two sets of plots. On the left, distribution maps (concentration profiles) are shown for each pure component as intensity heat maps. Colors range from dark blue (lowest intensity) to dark red (highest intensity). On the right, the spectral profiles for the corresponding pure components are shown, using the same colors as in the 'Pure Spectra Estimation (Initial Values)' plot (Step 19). If the 'Mark Purest' checkbox is checked (Step 21) before the 'Show Component Maps' button is pressed, the purest pixel for each component is also marked in their respective Component Map plot.

▲ CRITICAL STEP Component maps should match anatomical features of interest and should be in agreement with features that are observable in the visible image, the Total Intensity Plot or with a priori biological knowledge. Random maps or maps

that do not reflect the biological features of the sample are likely to be artifacts or products of a bad pre-processing. In such cases, a new model should be tested by changing key MCR parameters, such as the number of components (return to Step 18), noise in initial estimates (return to Step 20) or maximum number of iterations and convergence limits (return to Step 25; **Fig. 1**). Alternatively, pre-processing parameters can also be adjusted (return to Step 10). If none of the above parameter changes results in an improved match of the Component Maps to biological features, consultation with chemometrics and spectroscopy experts is necessary.
▲ CRITICAL STEP Spectra of the resolved pure components should be generally meaningful and never completely different from the features of the raw spectra. If only artifacts are resolved in ALL components, consultation with chemometrics and spectroscopy experts is necessary.
? TROUBLESHOOTING

32| (Optional) Save the 'Total Intensity Map' plot by clicking on the 'Save Intensity Map' button of the main interface. A save dialog box opens, with prefilled values for filename, format and location, which can be changed.
? TROUBLESHOOTING

33| Save the most important MCR-ALS results by clicking on the 'Save MCR Results' button. A save dialog box opens, with prefilled values for filename, format and location. Although filename and location can be changed, the format needs to remain .mat. The saved .mat file contains four variables. The variable *COpt* stores the optimized concentration profiles of the pure resolved components in $M \times N$ format, where M is the total number of pixels and N is the number of components (i.e., each column contains the concentration of a single pure component in all pixels; see C in **Fig. 5**). The variable *SOpt* stores the optimized spectral profiles of the pure resolved components in $N \times W$ format, where N is the number of components and W is the number of wavenumbers (i.e., each row represents one spectrum, see S^T in **Fig. 5**). The variable *R2Opt* contains the variance explained at the last iteration, expressed in values of percentage divided by 100 (i.e., *R2Opt* = 0.98 means 98% of variation explained). The variable *SDOpt* contains two numbers: the percentage lack of fit in terms of PCA and in terms of the experimental data.

(Optional) Reference spectra matching steps (evaluation) ● TIMING **~30 s–5 min**
34| Load reference spectra by clicking on the 'Load Reference Spectra' button in the 'Match Components to Reference Spectra' frame in the middle right of the interface. A load dialog box opens, in which multiple files can be selected, each file containing only a single reference spectrum.
▲ CRITICAL STEP Spectra must be in tab- or space-separated ASCII format (e.g., standard .txt files), with the first column containing wavenumbers and the second column containing the corresponding intensities. No other formats are accepted.
▲ CRITICAL STEP All reference spectra must be in the same folder, and they must have unique alphanumeric filenames (i.e., only unaccented Latin letters and numbers and underscore are allowed; special characters are not allowed). It is also important to have filenames that are representative for the reference compound, as these names will be displayed in legends and tables.
? TROUBLESHOOTING

35| (Optional) Tick the 'Pre-treat References' checkbox to pre-process the loaded reference spectra, using the same parameters as for image pre-processing (λ and P values for the baseline correction, Step 10, and polynomial order and frame size for the S-G smoothing, Step 12). Reference spectra and the pure resolved component spectra are always area normalized before reference matching, irrespective of the 'Pre-treat References' and 'Area Normalization' (Step 13) checkboxes. This is to ensure that total intensity differences do not affect reference matching, which aims to provide qualitative and not quantitative matches.

36| (Optional) Show the loaded reference spectra in a separate window by clicking on the 'Show Reference Spectra' button. This can be used to evaluate the quality of the reference spectra and to determine whether pre-processing is necessary or not. The reference spectra shown in this plot are always area normalized.
? TROUBLESHOOTING

37| Perform reference matching on the basis of Euclidean distances by clicking on the 'Perform Reference Matching (Dot Product)' button. The results are automatically displayed in the table below this button, showing the percentage match of each reference spectrum to each pure component. Depending on the number of components and reference spectra, scrolling may be needed to display all values.
▲ CRITICAL STEP The resulting percentage matches should always be considered as indicative only and should always be evaluated by visual inspection of the matches (Step 38 below). See the 'Reference spectra matching (chemical evaluation)' section of the INTRODUCTION regarding false-positive and false-negative matches. The percentage matches are determined for each compound individually and do not add up to 100% total.

38| Always inspect the matching results of Step 37 by clicking on the 'Show Matches' button. A new window opens, which contains one plot for each pure resolved component.
▲ CRITICAL STEP Critically examine matching results to exclude false-positive and false-negative matches (see the 'Reference spectra matching (chemical evaluation)' section in the INTRODUCTION). In case of uncertainties, consult a spectroscopy expert to avoid overinterpretation of the results.
! CAUTION In this plot, the pure component colors (Step 19) are not maintained. Instead, each component has its own separate plot in which it is displayed in solid thick black lines, whereas all reference spectra are displayed in dashed thin colored lines, in accordance with the plot's own legend.

39| Save the reference matching results by clicking the 'Save Match Results' button. A save dialog box opens, with prefilled values for filename, format and location. Although filename and location can be changed, the format needs to remain .mat. The saved .mat file contains the same table as displayed in the main interface, including row and column headings (i.e., component and reference names, respectively).

(Optional, recommended) Image segmentation steps (evaluation) ● TIMING **~10 s–5 min**
40| Determine the number of clusters for segmentation maps. This step can be performed using option A or B. Manual input (option B) is needed when segmentation maps with different numbers of clusters need to be tested (Step 41), or when the biological question at hand demands a certain number of clusters. In these cases, silhouette clustering (option A) is not used or its results need to be overruled.

(A) 'Silhouette Clusters' button

(i) Click on the 'Silhouette Clusters' button. This performs silhouette clustering to determine the number of clusters for segmentation maps. A message box appears to prompt the user to wait while silhouette clustering is in progress. When the process is completed, the message box automatically closes and the 'Number of clusters' text box updates. Although silhouette clustering is optional, it is recommended to get an initial overview of the data.

(B) Manual input

(i) Enter the number of clusters directly in the 'Number of clusters' text box.

41| Click on the 'K-Means Clustering' button to perform k-means clustering using the number of clusters determined in Step 40. This automatically updates the 'Segmentation Map (Clusters)' and 'Centroid Profiles' plots. A legend is automatically created for the 'Segmentation Map (Clusters)' plot to show the colors representing each cluster. The colors used in the 'Centroid Profiles' plot refer to the resolved pure components, and they are the same throughout the interface. Therefore, the 'Centroid Profiles' plot has no separate legend. Instead, the legend of the 'Pure Spectra Estimation (Initial Values)' plot (Step 19) should be consulted.
▲ CRITICAL STEP If segmentation maps are expected to provide detailed information regarding chemically distinct zones in the sample, different numbers of clusters may need to be tested. In that case, return to Step 40, option B.
! CAUTION The cluster number to which a particular pixel in the image belongs is randomly determined by *k*-means clustering. Therefore, re-running *k*-means clustering can result in different coloring. However, the clustering results (boundaries of clusters) do not change unless a different number of clusters is selected. In other words, the same pixels will belong to the same cluster; only the coloring of the 'Segmentation Map (Clusters)' plot is altered, together with the 'Centroid Profiles' plot to reflect the change in cluster order.

42| (Optional) Save the 'Segmentation Map (Clusters)' plot and legend by clicking on the 'Save Segmentation Plot' button. A save dialog box opens, with prefilled values for filename, format and location, which can be changed.
? TROUBLESHOOTING

43| (Optional) Save the 'Centroid Profiles' plot by clicking on the 'Save Centroid Plot' button. A save dialog box opens, with prefilled values for filename, format and location, which can be changed.
? TROUBLESHOOTING

44| (Optional) Save the results of the *k*-means clustering by clicking on the 'Save Segment Results' button. A save dialog box opens, with prefilled values for filename, format and location. Although filename and location can be changed, the format needs to remain .mat. The saved .mat file contains the variables 'IDX', 'Centr' and 'AverageClusterSpectra'. The variable 'IDX' stores the cluster number to which each pixel of the image belongs. The variable 'Centr' is a $K \times N$ matrix (where K is the number of clusters and N is the number of components), which describes the contribution of each component to each cluster. The 'AverageClusterSpectra' is a $K \times (W + 1)$ matrix (where K is the number of clusters and W is the number of wavenumbers—i.e., spectral dimension), which contains the cluster number in the first column and the average spectrum of each cluster in the corresponding row.

Finishing steps ● TIMING ~10–20 s

45| Close the interface window to finish the analysis. The most important variables remain in the MATLAB basic workspace, and they can still be saved until Step 47.
! CAUTION Unsaved plots of the main interface window cannot be recovered after this step.

46| Close all additional open figure windows (Eigen Vectors, Pre-MCR PCA results, Component Maps, Reference Spectra and Reference Matches) individually.
! CAUTION Unsaved plots cannot be recovered after closing their respective figure windows.

47| Type 'clear all' at the MATLAB Command Window prompt to clear the MATLAB workspace and memory from all variables.
! CAUTION Unsaved data cannot be recovered after this step.

48| Type 'clc' at the MATLAB Command Window prompt to clear the MATLAB Command Window.

49| If new image data need to be processed, return to Step 4.

50| After processing the final data, close MATLAB. The next analysis will have to start from Step 1.

? TROUBLESHOOTING

Troubleshooting advice can be found in **Table 2**.

TABLE 2 | Troubleshooting table.

Step	Problem	Possible reason	Solution
4	No EDITOR tab in MATLAB	The 'Editor' panel is not active	Click on the filename in the 'Editor' panel, which will bring up the EDITOR tab
5, 9, 29, 34	No files are selectable	File extension does not match	Change the drop-down menu of the open dialog box to show all file types Note that MATLAB is case-sensitive (i.e., .jpg and .JPG are not equivalent)
		Finder does not update (Mac users only)	Change the drop-down menu of the open dialog box to show all file types and select the file even if it is grayed out
	File does not load	Improper file formatting or name	Make sure that the file is in the correct format (see the Input data section of this PROCEDURE), and that the folder names do not contain special characters
6	Script is extremely slow to start	Too large data files in .txt format	Save the data in .mat format
		Insufficient computing power	Run MATLAB in native environments and not through hosted environments, e.g., parallels and so on Shut down unnecessary processes

(continued)

TABLE 2 | Troubleshooting table (continued).

Step	Problem	Possible reason	Solution
7, 8, 27, 32, 42, 43	Saved plots are of suboptimal quality	File format is suboptimal	Change the file type in the save dialog box
	Text is cropped in the GUI	Low-resolution computer screen	If possible, increase the screen resolution. Maximize the GUI window. Refer to **Supplementary Figure 1** for the full version of texts in the GUI
10, 11, 19, **Box 1**, step 3	Legend obscures the 'Selected Spectrum' plot	High-intensity data points in the low wavenumber region of the spectrum	Click and drag on the legend to move it to another location
Box 1, step 6	No white light image is loaded at the start	No exact match of the data filename with lowercase .jpg extension for the figure was found	Manually load a white light image figure by the 'Load White Light Image' button in the 'Visualization' frame (Step 9 in the PROCEDURE)
12	The smoothed spectrum in the 'Selected Spectrum' plot does not update after changing S-G smoothing parameters	Variables are not updated in the MATLAB script	Press 'Enter'/'Return' after typing the value in the 'Order' and 'Frame' textboxes. Unmark and mark the 'S-G filtering' checkbox
18	Changing the number of components does not refresh the 'Pure Spectra Estimates (Initial Values)'	Variables are not cleared and updated in the MATLAB script, especially when changing to lower number of components	Restart the GUI and the entire analysis by completing Steps 44–47 and by returning to Step 4. CAUTION This will result in the loss of all unsaved data
20	The 'Pure Spectra Estimation (Initial Values)' plot does not refresh after changing the noise level	Variables are not updated in the MATLAB script	Force update by changing the number of components in Step 18 to a lower value and then back to the desired value
21, 30	Changing the number of components does not refresh the Pure Pixel markings in the 'Total Intensity' plot and in the 'Component Maps'	Variables are not cleared and updated in the MATLAB script, especially when changing to lower number of components	Force updating the locations of the purest pixels, untick and tick the 'Mark Purest' checkbox. Please note that this only works if a higher number of components are selected than previously. In case a lower number of components are selected, the 'Mark Purest' checkbox is permanently unable to properly display the fewer components and the interface needs to be restarted by completing Steps 44–47 and returning to Step 4. Note that his will result in the complete loss of unsaved data
26	MCR-ALS text results are illegible or cropped in the GUI	Low-resolution computer screen	Increase the screen resolution Maximize the GUI window Read the MCR-ALS results in full text version in the MATLAB Command window
30, 31	The separate window containing the pure component maps does not show/hide the purest pixel after ticking/unticking the 'Mark Purest' checkbox of the main GUI window	Once the separate figure window is open, it does not refresh automatically	First close the separate window showing the component maps, tick or untick the 'Mark Purest' checkbox on the main interface window and then click on the 'Show Component Maps' button to reopen the updated component maps figure window
36	The separate window containing the reference spectra does not update when ticking or unticking the 'Pre-treat References' checkbox in the main interface window	Once the separate figure window is open, it does not refresh automatically	First close the separate window showing the reference spectra plot, tick or untick the 'Pre-treat References' checkbox on the main interface window and then click on the 'Show Reference Spectra' button to reopen the updated reference spectra plot window

● TIMING

Steps 1–9, starting steps: ~30 s–5 min
Steps 10–16, pre-processing steps: ~1–20 min
Steps 17–28, data analysis steps: ~2–30 min
Steps 29–33, visualization steps: ~10 s–1 min
Steps 34–39, reference spectra matching steps: ~30 s–5 min
Steps 40–44, image segmentation steps: ~10 s–5 min
Steps 45–50, finishing steps: ~10–20 s

ANTICIPATED RESULTS

Although the results described below are based on previous works by the authors[24,31,34,40], they were not included in these original publications and were specifically selected to demonstrate key features of the present protocol, including a deliberately bad example to illustrate limitations (**Fig. 6**, left). By using vibrational microspectroscopic (hyperspectral) images of a biological sample, the described procedure is able to identify (i) the number of distinct chemical components that can be differentiated in the sample on the basis of their spectral profiles; (ii) the pure spectral profiles of each component; (iii) the relative concentration of each component in each pixel of the image (plotted as a distribution map); and (iv) the number and distribution of zones with distinct chemical characteristics (segmentation). In addition, the spectral profiles of the resolved unique chemical components can be matched to the spectral profiles of reference compounds for the purposes of identification. Below, we illustrate the power of the method by finding chemically distinct zones within cell wall layers of woody plant tissues (tension wood of hybrid aspen, **Fig. 6**, right) and within a larger mammalian tissue sample (mouse pancreas, **Fig. 2**) in the absence of clear visible boundaries.

Resolving cell wall layers

As visible features often do not exist for differentiating the cell wall layers, and simple distance measurements from easily recognized features (such as the lumen) cannot be used owing to cell wall thickness variations, spectral information needs to be used for this purpose. Heat maps (based either on single-band intensity or on entire spectral profiles) are of limited use in this respect, as they are often unspecific or only provide concentration gradients, but they do not have distinct boundaries among zones (**Fig. 6d**, right). By using our protocol, however, cell wall layers can be clearly defined in segmentation schemes.

In the example shown in **Figure 6**, the right column is the hyperspectral image of the cross-section of a hybrid aspen stem containing tension wood fibers. We chose this example for demonstration owing to clear and characteristic chemical composition differences of tension wood fibers, allowing easy validation of the results. In addition to the cell wall layers of normal wood fibers[51], tension wood fibers have a thick cell wall layer on the lumen side of the cell wall. This additional layer (G-layer) is extremely rich in cellulose, but it is mostly free from the lignin that is present in the underlying layers (Felten and Sundberg[52], and references therein). As cellulose and lignin have clearly distinguishable Raman spectroscopic fingerprints, and as the G-layer is rather thick, it is expected to be easily distinguishable from the underlying cell wall layers. Accordingly, the pure components clearly resolve both lignin and cellulose spectral profiles, and the corresponding distribution maps highlight areas that are rich in each (**Fig. 6d,e**, right; C1 and C2 spectral profiles and maps for lignin and cellulose, respectively). In particular, the distribution map for component 2 (C2) can be used to select clear G-layer pixels in the image. However, to separate the underlying, thinner and chemically more similar cell wall layers, further analysis (image segmentation) is required.

Four clusters were selected for *k*-means clustering on the basis of Silhouette values and a priori knowledge of the sample, as we expected to resolve the lumen and three cell wall layers: (i) the cellulose-rich G-layer on the lumen side; (ii) the lignin-rich middle lamella separating adjacent cells; and (iii) the underlying S-layer, which is a mixture of different thin and chemically similar layers that cannot be distinguished at the spatial resolution of the experiment (see below). The segmentation map (**Fig. 6f**, right, segmentation map) clearly defines these zones. Cluster 2 (light blue) is the G-layer, and it has the highest contribution of component 2 (cellulose, green contribution in the Centroid Profile plot, **Fig. 6g**, right). Cluster 1 (dark blue) represents the S-layer with high contributions from both lignin (component 1, blue line) and cellulose (component 2, green line). Cluster 4 (brown) is almost exclusively lignin (component 1, blue line), and it can thus be identified as the middle lamella. Anatomically, the dark blue zone on the right side of the segmentation map, wedged between two G-layers (light blue), should be expected to contain pixels describing the middle lamella, and thus it should have a brown line in the center (indicating the presence of cluster 4).

There are two main reasons for the absence of this line. First, the lateral resolution of the image is too low to resolve zones thinner than 1 μm. Second, even if the middle lamella is almost 1 μm thick, it may not coincide with the pixel boundaries; i.e., parts of this cell wall layer can belong to different pixels. This decreases its contribution to the spectrum of those pixels, which have significant signals from neighboring zones. In short, no clear pixel for the middle lamella can be found in this part of the image, which is also slightly out of focus, further decreasing detection limits of smaller contributions.

Nevertheless, the middle lamella could be detected in the cell walls that appear to the left in the image, as it was better resolved in that location: it was thicker, owing to the cell corner being covered in this part of the image, it coincided more with pixel boundaries and was more in focus. Finally, cluster 3 (yellow) represents the lumen, which has the highest contributions from components 3 and 4, dominated by noise (red and cyan, respectively in **Fig. 6e**, right column, C3 and C4). On the basis of these results, representative pixels for not only the G-layer but also for other cell wall layers can be identified, and their representative spectra can be extracted and compared with spectra of the corresponding cell wall layers of other genotypes or trees growing in the absence of tension wood formation. This example clearly illustrates the power of MCR-ALS in component identification, both for spectral profiles and for component distribution maps. It also highlights how component maps may not be sufficient for the identification of chemically distinct zones with clear boundaries. In those cases, segmentation maps help achieve clear results.

It is important to note that the clear interpretation of the clusters is only possible because the image segmentation is based on the resolved concentration profiles by MCR-ALS. This not only speeds up the segmentation process compared with classical approaches using the full pixel spectra but also allows the straightforward interpretation of centroid profiles as mixtures of pure components in different proportions.

Resolving chemically distinct zones in a mammalian tissue

The progression of type I diabetes has been monitored by following the loss of pancreatic beta cells using 3D optical tomography[53–55], as these beta cells are directly responsible for insulin production. Beta cells are organized into the islets of Langerhans, which are scattered throughout the exocrine tissue. However, visually differentiating these islets in the surrounding exocrine tissue without staining is difficult, and thus they serve as an ideal example to demonstrate the biomedical applicability of hyperspectral imaging using the present procedure.

By following the steps of the present protocol, the FTIR microspectroscopic image of a mouse pancreas section was resolved into five pure components (**Fig. 2**). None of the pure components are pure chemical compounds, but they have contributions from various proteins (bands ~1,550 and 1,650 cm^{-1}), as well as from lipids and carbohydrates (bands between 1,000 and 1,200 cm^{-1}), and the embedding medium (mostly in C2 and C5, green and violet, respectively, and to a lesser extent in C1 (blue) owing to smearing during sectioning, **Fig. 2c**, spectral profiles). Although it can be important to know the exact chemical composition of different zones within the tissue (or even within cells, as in the cell wall layer example above), here we exclusively focus on using our procedure to detect a known feature in the image (the islet cells embedded in the exocrine tissue), on the basis of its unique spectral fingerprint. Thus, interpreting the spectra in terms of exact chemical compounds is beyond the scope of the protocol, and the emphasis is on the different spatial maps. The pure component maps (**Fig. 2c**, concentration profiles) are unable to detect islet cells directly, but they already highlight problematic areas (see the component maps of C2 and C5 in particular, but also of C1 to a lesser extent, as according to their spectral profiles the embedding medium is clearly contributing to the signal in these zones; **Fig. 2c**). This is also clear in the segmentation map (cluster 3, brown, **Fig. 2d**). However, the segmentation map also resolves chemically different zones within the pancreatic tissue, with cluster 1 (dark blue) associated with the islet cells and cluster 2 (light green) associated with the exocrine tissue. Note also that these zones have the least contribution from embedding medium (contributions of C2 and C5, green and violet, respectively, in the Centroid Plot; **Fig. 2e**), and thus they provide the most reliable spectral profiles of the tissue.

Extracting spectra from these chemically different zones at different stages of the disease development provides valuable information regarding the biochemical changes associated with the pathological conditions. In general, such information can be used for assessing risk factors, for monitoring disease progression or even for developing targeted therapies for selected cell types or tissue zones.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS We thank the Vibrational Spectroscopy Core Facility of the Chemical Biological Centre at Umeå University for full Raman and FTIR microspectroscopy instrumentation access and resources dedicated to method development. We thank B. Sundberg at the Umeå Plant Science Centre for initializing the project using hybrid aspen and continued support. U. Ahlgren and C. Nord at the Umeå Centre for Molecular Medicine are acknowledged for providing the mouse pancreas sample. A. de Juan and R. Tauler acknowledge financial support of the European Union project CHEMAGEB (Chemometrics and high-throughput omics analytical methods for assessment of global change effects on environmental and biological systems) and Joaquim Jaumot from the Spanish government (grant CTQ2012-11572). J. Felten and H. Hall acknowledge support by grants from the Swedish Research Council, Formas, Vinnova and Bio4Energy, the Swedish Programme for renewable energy.

AUTHOR CONTRIBUTIONS The project was initiated by J.F. and A.G.; plant samples were prepared by J.F.; Raman hyperspectral images were recorded by J.F. and A.G.; FTIR hyperspectral image were recorded by A.G.; Raman reference compounds were prepared and recorded by J.F. and A.G.; MCR-ALS was developed by R.T., J.J. and A.de J.; the present MATLAB GUI was designed, written and tested by A.G., J.F., H.H. and J.J.; the manuscript was written by J.F., H.H., R.T., A.de J. and A.G.; and the project was coordinated and supervised by A.G.

COMPETING FINANCIAL INTERESTS The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Geladi, P., Grahn, H. & Burger, J. in *Techniques and Applications of Hyperspectral Image Analysis* (eds. Grahn, H.F. & Geladi, P.L.M.) (John Wiley & Sons, 2007).

2. Hall, H., Cheung, J. & Ellis, B. Immunoprofiling reveals unique cell-specific patterns of wall epitopes in the expanding *Arabidopsis* stem. *Plant J.* **74**, 134–147 (2013).
3. Wilson, S. & Bacic, A. Preparation of plant cells for transmission electron microscopy to optimize immunogold labeling of carbohydrate and protein epitopes. *Nat. Protoc.* **7**, 1716–1727 (2012).
4. Fabian, H. *et al.* Diagnosing benign and malignant lesions in breast tissue sections by using IR-microspectroscopy. *Biochim. Biophys. Acta* **1758**, 874–882 (2006).
5. Nijssen, A. *et al.* Discriminating basal cell carcinoma from its surrounding tissue by Raman spectroscopy. *J. Invest. Dermatol.* **119**, 64–69 (2002).
6. Sobottka, S., Geiger, K., Salzer, R., Schackert, G. & Krafft, C. Suitability of infrared spectroscopic imaging as an intraoperative tool in cerebral glioma surgery. *Anal. Bioanal. Chem.* **393**, 187–195 (2009).
7. Gorzsás, A., Stenlund, H., Persson, P., Trygg, J. & Sundberg, B. Cell-specific chemotyping and multivariate imaging by combined FT-IR microspectroscopy and orthogonal projections to latent structures (OPLS) analysis reveals the chemical landscape of secondary xylem. *Plant J.* **66**, 903–914 (2011).
8. Gierlinger, N. & Schwanninger, M. Chemical imaging of poplar wood cell walls by confocal Raman microscopy. *Plant Physiol.* **140**, 1246–1254 (2006).
9. Chang, S.-S., Salmén, L., Olsson, A.-M. & Clair, B. Deposition and organisation of cell wall polymers during maturation of poplar tension wood by FTIR microspectroscopy. *Planta* **239**, 243–254 (2013).
10. Pesquet, E. *et al.* Non-cell-autonomous postmortem lignification of tracheary elements in *Zinnia elegans*. *The Plant Cell* **25**, 1314–1328 (2013).
11. Tsai, A. *et al.* Constitutive expression of a fungal glucuronoyl esterase in *Arabidopsis* reveals altered cell wall composition and structure. *Plant Biotechnol. J.* **10**, 1077–1087 (2012).
12. Horvath, L. *et al.* Distribution of wood polymers within the cell wall of transgenic aspen imaged by Raman microscopy. *Holzforschung* **66**, 717–725 (2012).
13. Schmidt, M. *et al.* Label-free *in situ* imaging of lignification in the cell wall of low lignin transgenic *Populus trichocarpa*. *Planta* **230**, 589–597 (2009).
14. Gierlinger, N., Keplinger, T. & Harrington, M. Imaging of plant cell walls by confocal Raman microscopy. *Nat. Protoc.* **7**, 1694–1708 (2012).
15. Richter, S., Müssig, J. & Gierlinger, N. Functional plant cell wall design revealed by the Raman imaging approach. *Planta* **233**, 763–772 (2011).
16. Gierlinger, N. *et al.* Cellulose microfibril orientation of *Picea abies* and its variability at the micron-level determined by Raman imaging. *J. Exp. Bot.* **61**, 587–595 (2010).
17. Gierlinger, N., Schwanninger, M., Reinecke, A. & Burgert, I. Molecular changes during tensile deformation of single wood fibers followed by Raman microscopy. *Biomacromolecules* **7**, 2077–2081 (2006).
18. Naumann, A., Navarro-Gonzalez, M., Peddireddi, S., Kues, U. & Polle, A. Fourier transform infrared microscopy and imaging: detection of fungi in wood. *Fungal Genet. Biol.* **42**, 829–835 (2005).
19. Wilson, R.H. *et al.* The mechanical properties and molecular dynamics of plant cell wall polysaccharides studied by Fourier-transform infrared spectroscopy. *Plant Physiol.* **124**, 397–405 (2000).
20. Faix, O. Classification of lignins from different botanical origins by FT-IR spectroscopy. *Holzforschung* **45**, 21–27 (1991).
21. Kataoka, Y. & Kondo, T. Quantitative analysis for the cellulose Iα crystalline phase in developing wood cell walls. *Int. J. Biol. Macromol.* **24**, 37–41 (1999).
22. Akerholm, M., Hinterstoisser, B. & Salmen, L. Characterization of the crystalline structure of cellulose using static and dynamic FT-IR spectroscopy. *Carbohydr. Res.* **339**, 569–578 (2004).
23. Wetzel, D. in *Infrared and Raman Spectroscopic Imaging* (eds. Salzer, R. & Siesler, H.W.) (Wiley-VCH, 2009).
24. Gorzsás, A. & Sundberg, B. Chemical fingerprinting of *Arabidopsis* using Fourier transform infrared (FT-IR) spectroscopic approaches. *Methods Mol. Biol.* **1062**, 317–352 (2014).
25. Jirasek, A., Schulze, G., Yu, M.M.L., Blades, M.W. & Turner, R.F.B. Accuracy and precision of manual baseline determination. *Appl. Spectrosc.* **58**, 1488–1499 (2004).
26. Eilers, P.H.C. Parametric time warping. *Anal. Chem.* **76**, 404–411 (2004).
27. Savitzky, A. & Golay, M.J.E. Smoothing + differentiation of data by simplified least squares procedures. *Anal. Chem.* **36**, 1627 (1964).
28. MathWorks. MATLAB http://www.mathworks.com/help/matlab/index.html (2013).
29. Stenlund, H., Gorzsás, A., Persson, P., Sundberg, B. & Trygg, J. Orthogonal projections to latent structures discriminant analysis modeling on in situ FT-IR spectral imaging of liver tissue for identifying sources of variability. *Anal. Chem.* **80**, 6898–6906 (2008).
30. Baranska, M., Schulz, H., Rosch, P., Strehle, M.A. & Popp, J. Identification of secondary metabolites in medicinal and spice plants by NIR-FT-Raman microspectroscopic mapping. *Analyst* **129**, 926–930 (2004).
31. de Juan, A., Maeder, M., Hancewicz, T., Duponchel, L. & Tauler, R. in *Infrared and Raman Spectroscopic Imaging* (eds. Salzer, R. & Siesler, H.W.) Ch. 2, 65–106 (Wiley-VCH, 2009).
32. Bonnier, F. & Byrne, H.J. Understanding the molecular information contained in principal component analysis of vibrational spectra of biological systems. *Analyst* **137**, 322–332 (2012).
33. Tran, T.N., Wehrens, R. & Buydens, L.M.C. Clustering multispectral images: a tutorial. *Chemometrics Intellig. Lab. Syst.* **77**, 3–17 (2005).
34. Piqueras, S., Duponchel, L., Tauler, R. & de Juan, A. Resolution and segmentation of hyperspectral biomedical images by multivariate curve resolution-alternating least squares. *Anal. Chim. Acta* **705**, 182–192 (2011).
35. Nascimento, J.M.P. & Bioucas-Dias, J.M. Vertex component analysis: a fast alogrithm to unmix hyperspectral data. *IEEE Trans. Geosci. Remote Sens.* **43**, 898–910 (2005).
36. Bioucas-Dias, J.M. *et al.* Hyperspectral unmixing overview: geometrical, statistical, and sparse regression-based approaches. *IEEEE J. Stars* **5**, 354–379 (2012).
37. Krafft, C. *et al.* Crisp and soft multivariate methods visualize individual cell nuclei in Raman images of liver tissue sections. *Vib. Spectrosc.* **55**, 90–100 (2011).
38. Piqueras, S., Burger, J., Tauler, R. & de Juan, A. Relevant aspects of quantification and sample heterogeneity in hyperspectral image resolution. *Chemometrics Intellig. Lab. Syst.* **117**, 169–182 (2012).
39. Zhang, X. & Tauler, R. Application of multivariate curve resolution alternating least squares (MCR-ALS) to remote sensing hyperspectral imaging. *Anal. Chim. Acta* **762**, 25–38 (2013).
40. Jaumot, J., Gargallo, R., de Juan, A. & Tauler, R. A graphical user-friendly interface for MCR-ALS: a new tool for multivariate curve resolution in MATLAB. *Chemometrics Intellig. Lab. Syst.* **76**, 101–110 (2005).
41. de Juan, A., Rutan, S.C. & Tauler, R. in *Comprehensive Chemometrics* (eds. Brown, S., Tauler, R. & Walczak, R.) 325–344 (Elsevier B. V., 2009).
42. Tauler, R., Smilde, A. & Kowalski, B. Selectivity, local rank, 3-way data-analysis and ambiguity in multivariate curve resolution. *J. Chemom.* **9**, 31–58 (1995).
43. Windig, W. & Guilment, J. Interactive self-modeling mixture analysis. *Anal. Chem.* **63**, 1425–1432 (1991).
44. Windig, W. Spectral data files for self-modeling curve resolution with examples using the Simplisma approach. *Chemometrics Intellig. Lab. Syst.* **36**, 3–16 (1997).
45. Batonneau, Y., Laureyns, J., Merlin, J.C. & Bremard, C. Self-modeling mixture analysis of Raman microspectrometric investigations of dust emitted by lead and zinc smelters. *Anal. Chim. Acta* **446**, 23–37 (2001).
46. de Juan, A., Maeder, M., Hancewicz, T. & Tauler, R. Local rank analysis for exploratory spectroscopic image analysis. Fixed size image window-evolving factor analysis. *Chemometrics Intellig. Lab. Syst.* **77**, 64–74 (2005).
47. de Juan, A., Maeder, M., Hancewicz, T. & Tauler, R. Use of local rank-based spatial information for resolution of spectroscopic images. *J. Chemom.* **22**, 291–298 (2008).
48. Li, J.F., Hibbert, D.B., Fuller, S., Cattle, J. & Way, C.P. Comparison of spectra using a Bayesian approach. An argument using oil spills as an example. *Anal. Chem.* **77**, 639–644 (2005).
49. Mark, H. & Workman, J. *Chemometrics in Spectroscopy* (Elsevier, 2007).
50. Linusson, A., Wold, S. & Norden, B. Fuzzy clustering of 627 alcohols, guided by a strategy for cluster analysis of chemical compounds for combinatorial chemistry. *Chemometrics Intellig. Lab. Syst.* **44**, 213–227 (1998).
51. Plomion, C., Leprovost, G. & Stokes, A. Wood formation in trees. *Plant Physiol.* **127**, 1513–1523 (2001).
52. Felten, J. & Sundberg, B. in *Cellular Aspects of Wood Formation Plant Cell Monographs* (ed. Fromm, J.) 203–224 (Springer, 2013).
53. Alanentalo, T. *et al.* Tomographic molecular imaging and 3D quantification within adult mouse organs. *Nat. Methods* **4**, 31–33 (2007).
54. Alanentalo, T. *et al.* Quantification and three-dimensional imaging of the insulitis-induced destruction of beta cells in murine type 1 diabetes. *Diabetes* **59**, 1756–1764 (2010).
55. Hornblad, A., Cheddad, A. & Ahlgren, U. An improved protocol for optical projection tomography imaging reveals lobular heterogeneities in pancreatic islet and beta cell mass distribution. *Islets* **3**, 204–208 (2011).