Bioinformatics, 31, 2015, i375–i384
doi: 10.1093/bioinformatics/btv266
ISMB/ECCB 2015

OXFORD

Using collective expert judgements to evaluate quality measures of mass spectrometry images

Andrew Palmer[1,2], Ekaterina Ovchinnikova[1,3], Mikael Thuné[4], Régis Lavigne[5], Blandine Guével[5], Andrey Dyatlov[2,6], Olga Vitek[7], Charles Pineau[5], Mats Borén[4], Theodore Alexandrov[1,2,6,8,*]

[1]European Molecular Biology Laboratory, Heidelberg, Germany, [2]Center for Industrial Mathematics, University of Bremen, Bremen, Germany, [3]High Performance Humanoid Technologies Lab, Institute for Anthropomatics, Karlsruhe Institute of Technology, Karlsruhe, Germany, [4]Denator, Uppsala, Sweden, [5]Protim, Inserm U1085 - Irset, University of Rennes 1, Rennes, France, [6]SCiLS GmbH, Bremen, Germany, [7]College of Computer and Information Science, Northeastern University, Boston, MA, USA and [8]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA, USA

*To whom correspondence should be addressed.

Abstract

Motivation: Imaging mass spectrometry (IMS) is a maturating technique of molecular imaging. Confidence in the reproducible quality of IMS data is essential for its integration into routine use. However, the predominant method for assessing quality is visual examination, a time consuming, unstandardized and non-scalable approach. So far, the problem of assessing the quality has only been marginally addressed and existing measures do not account for the spatial information of IMS data. Importantly, no approach exists for unbiased evaluation of potential quality measures.

Results: We propose a novel approach for evaluating potential measures by creating a gold-standard set using collective expert judgements upon which we evaluated image-based measures. To produce a gold standard, we engaged 80 IMS experts, each to rate the relative quality between 52 pairs of ion images from MALDI-TOF IMS datasets of rat brain coronal sections. Experts' optional feedback on their expertise, the task and the survey showed that (i) they had diverse backgrounds and sufficient expertise, (ii) the task was properly understood, and (iii) the survey was comprehensible. A moderate inter-rater agreement was achieved with Krippendorff's alpha of 0.5. A gold-standard set of 634 pairs of images with accompanying ratings was constructed and showed a high agreement of 0.85. Eight families of potential measures with a range of parameters and statistical descriptors, giving 143 in total, were evaluated. Both signal-to-noise and spatial chaos-based measures performed highly with a correlation of 0.7 to 0.9 with the gold standard ratings. Moreover, we showed that a composite measure with the linear coefficients (trained on the gold standard with regularized least squares optimization and lasso) showed a strong linear correlation of 0.94 and an accuracy of 0.98 in predicting which image in a pair was of higher quality.

Availability and implementation: The anonymized data collected from the survey and the Matlab source code for data processing can be found at: https://github.com/alexandrovteam/IMS_quality.

Contact: theodore.alexandrov@embl.de

1 Introduction

1.1 Motivation

1.1.1 The need for quality measures in IMS

IMS recently emerged as an analytical chemistry technique for untargeted and label-free molecular imaging of tissue sections, agar plates and cell cultures that can localize hundreds of molecules simultaneously with a high molecular specificity and sensitivity, and with cellular spatial resolution (Spengler, 2015). IMS was introduced for biological analysis in 1997 and rapidly commercialized so that it now provides a capable tool for molecular imaging directly from tissue sections with the increasing potential for routine

application. Matrix assisted laser desorption ionization (MALDI) is probably the most commonly encountered ionization technique used for IMS. IMS produces a dataset which can be viewed as a collection of mass spectra acquired at pixels of the analyzed area, or as a hyperspectral imaging dataset with thousands to millions of channels (ion images), each representing the spatial distribution of an ion with a particular mass to charge ratio (*mz*) value. The quality of the data and in particular of ion images produced is a key concern as this directly impacts the ability of a researcher to draw conclusions from the data (Deininger *et al.*, 2011). IMS is becoming a mature technique with standardized protocols (Chaurand, 2012; Schuerenberg and Deininger, 2010); but decisions on several experimental settings need to be made for each experiment, including the washing procedure, matrix application (in the case of MALDI) and acquisition parameters, with the aim to maximize the quality.

Currently, assessment of the quality of IMS data is performed by a mass spectrometry expert using their knowledge and intuition to evaluate the data. This includes visual examination of individual spectra and ion images. However, as IMS becomes faster and widespread, the data generation rate increases encompassing studies of cohorts (Balluff *et al.*, 2015) and hundreds of serial sections for 3D images (Oetjen *et al.*, 2013; Trede *et al.*, 2012) that makes it impractical for an expert to evaluate the data produced. Moreover, the optimization of experimental settings with the increase of considered parameters leads to the explosion of possible combinations of settings and, correspondingly, of datasets to be generated and evaluated.

A quantitative measure for objective, unbiased and automated evaluation of the quality of IMS data would enable optimizing experimental steps, continuous monitoring of instrument performance, evaluating the suitability of data for analysis and reporting results across instruments and laboratories. In this study, we propose using image-based quality measures and address the key question of their evaluation through creating a gold-standard dataset.

1.1.2 Creating gold standard using collective expert knowledge

For knowledge-intensive tasks, it is often the expert judgements that is the most important measure of quality of results, against which automatic methods can be evaluated. Moreover, it is assumed that collective judgments sourced from a crowd of experts are more reliable than judgement of any single expert (Gwet, 2012). Experts-annotated ground truth datasets are called 'gold-standard' and are used to train and test algorithms in a wide range of domains including natural language processing (Carletta, 1996), content analysis (Krippendorff, 2012) and clinical decision support systems (Berner, 2003). However, creating an annotated gold-standard data by involving experts is a challenging task. Particular attention should be paid to the experimental design and inter-annotator reliability (Gwet, 2012). In this study, we propose an experimental design and an open-source online survey platform for collecting, evaluating and compiling expert judgements of IMS data into a re-usable and transparent gold standard.

1.2 Related work

1.2.1 Collective expert judgement and crowdsourcing in mass spectrometry

Collective expert judgement for creating a gold-standard has not yet been introduced into the field of mass spectrometry or IMS and crowdsourcing has only been minimally introduced. In this section, we provide an overview of existing projects using crowdsourcing that can be considered as a large-scale collective expert judgement.

One example is MSiMass, an online list of *mz*-values reported in the IMS literature; containing 293 *mz*-values as of March 2015 (McDonnell *et al.*, 2014). A related initiative was proposed by (Römpp *et al.*, 2014) who implemented and reported submission of IMS datasets into the open PRIDE repository; containing one IMS dataset as of March 2015.

Use of crowdsourcing for non-IMS has a longer history. The most notable examples are the user-populated spectral repositories such as MassBank, HMDB, GOLM, mzCloud, Lifeline-S.O.S., ProteomicsDB, MetaboLights and GnPS. The Spectral Game is a web-based game with purpose where users are asked to match molecular structures with, in particular, mass spectra (Bradley *et al.*, 2009). A related initiative was reported by Du *et al.* (2014) who carried out a citizen-science project on natural products discovery by crowdsourcing the collection of soil samples which were later analyzed with mass spectrometry.

1.2.2 Assessing data quality in mass spectrometry

1.2.2.1 control and suitability tests. A broad topic in mass spectrometry where the quality of mass spectra is considered is the quality control which aims to detect changes in the instrumentation. For this, intra-experimental monitoring is performed by analyzing the same sample over weeks and months so that changes in the instrument performance can be seen (Abdel-Rehim, 2004). Similar to quality control, the system suitability tests serve to check that the instrumentation satisfies the specifications. For this, analysis of standardized mixtures is performed to benchmark the performance of a mass spectrometer (Dresen *et al.*, 2010). By examining the properties of spectra (e.g. peak shape or intensity), defects can be automatically detected (Mutton *et al.*, 2011).

1.2.2.2 Quality of tandem mass spectra. A particular problem needing estimation of quality of individual mass spectra is the matching of tandem mass spectra against databases. Here, the filtering out of low-quality spectra that are unlikely to return a database match can increases the matching efficiency. Heuristic approaches have been developed which typically examine the number of peaks and relative intensities of peaks detected (Bern *et al.*, 2004; Ma *et al.*, 2003). Machine learning approaches have been used to recognize 'good' spectra by training a support vector machine on sets of labelled 'good' and 'bad' spectra (Bern *et al.*, 2004). Typically this labelling is achieved by running database matching on a subset of the spectra with 'good' being manually annotated or those for which a match was found and 'bad' being those that were not matched (Ma *et al.*, 2003; Nesvizhskii *et al.*, 2006).

1.2.2.3 Quality of IMS data. Currently, no approach exist that could automatically assess the quality of IMS data and would account for specific properties of IMS data which are: complex spectra representing mixtures of analytes analyzed without separation, limited-to-no capabilities for tandem mass spectrometry, repetition of acquisition of spectra coming from a region of similar chemical composition and spatial imaging information.

Karlsson *et al.* (2014) recently reported on optimizing the experimental settings for MALDI IMS through using a variance-based measure of quality of spectra. Considering three measures of variance aided in minimizing the unwanted variance in MALDI IMS and to discover subtle changes in protein expression in various subregions of the brain. Note that in (Karlsson *et al.*, 2014) the imaging content was not considered. We briefly discussed the need for quality estimators in (Watrous *et al.*, 2011) and proposed a spectra-based quality test for MALDI time-of-flight IMS data.

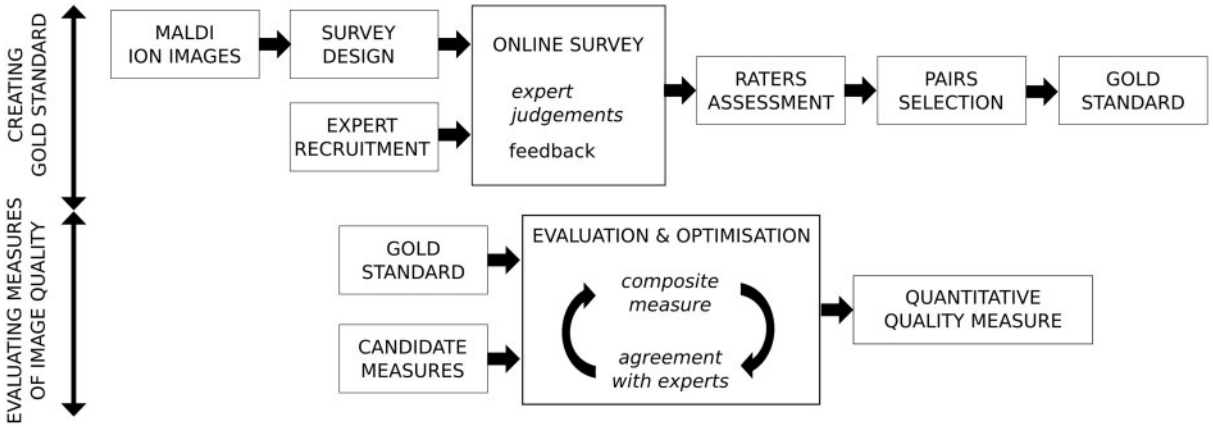

Fig. 1. The workflow of our study containing two parts: (i) creation of the gold-standard set of ion image pairs annotated by experts with relative quality of images in a pair and (ii) evaluating candidate measures of quality of ion images. The anonymized data from the survey and the Matlab source code for data analysis is available in the GitHub project repository

1.3 Problem statement and our approach

The quality of IMS data is currently assessed either by a subjective expert judgement or by examining individual spectra. In order to make this assessment automatic and at the same time to account for imaging information inherently present in IMS data, image-based quality measures are required.

Our approach to solve this problem has two essential components (see Fig. 1). First, propose to create a gold-standard set by involving the IMS community and recruiting a crowd of experts to provide their subjective estimation of the relative quality between a pair of images. Second, we propose to consider measures of quality of IMS data based on the image analysis principles, and to assess them for their ability to reproduce the human judgements.

The contribution of this article is the proposed approach for which mass spectrometry data was acquired, expert data ratings collected, a gold-standard set of pairs of images generated with relative quality ratings and new measures of quality of ion images evaluated on the gold-standard data.

2 Methods

2.1 Creating a gold standard through expert rating

We aimed at creating a gold-standard dataset consisting of pairs of ion images annotated with their relative quality that can be used later to evaluate candidate image-based measures of quality. This involved selection of ion images from MALDI-TOF IMS datasets, recruitment of experts to provide quality ratings through an online survey, assessment of the expert-provided ratings and compilation of the gold-standard.

2.1.1 Selection of ion images

2.1.1.1 MALDI IMS data. A 12 μm thick cryo-sections of wild-type mouse brain were prepared on conductive-coated slides that then underwent a washing procedure with either ethanol or propanol. Sinapinic acid matrix was applied with the ImagePrep (Bruker Daltonics, Bremen, Germany) either using the manufacturer's standard protocol or with an extended incubation time. MALDI IMS datasets were collected on an UltraFlex (Bruker Daltonics, Bremen, Germany) in linear positive-ion mode and in the mass range 2–20 kDa. A 75-μm raster size was selected and each spectrum was the sum of 100 or 600 laser shots at 2000 Hz repetition rate using either a tightly focused laser beam (with 5% random walk) or a static defocused beam that covered an equivalent area. In total 10 datasets were collected with different combinations of instrument parameters.

2.1.1.2 Selection of ion images for rating. We aimed to select 52 different and representative ion images to be used in the survey. (For explanations on why 52 images were needed, please see Section 2.1.2.) In order to avoid biasing the images selected with the authors' own opinions, a semi-automated selection procedure was used: For each dataset, peaks in the mean spectrum were determined from gradient turning points as per Coombes *et al.* (2005). The 10 most intense peaks were taken from each dataset and merged into a single aggregate list for all datasets, with any values within 1 Da averaged. Ion images were produced from each dataset for every peak in the aggregate list with a summation window of ± 5 Da around the peak centroid. The intensities of each image I were scaled to [0, 1]. The images were partitioned into 52 groups using k-means clustering with the Euclidean distance. For each cluster, an image with the closest distance to the cluster-average was selected.

2.1.2 Design of the online survey

For the selected ion images, our aim was to obtain relative ratings of their quality by involving experts in IMS. Our expectations were that although there is no precise formulation for the quality of ion images, experts use this concept in their everyday work and so they can assess whether images are of a high or low quality. To capture this, we generated all possible pairs of selected images, showed them to experts pair-by-pair, and asked the experts to provide a relative rating of quality between two images in a pair.

2.1.2.1 Interface design. A web-based survey was deployed with three sections: an introduction with instructions; a series of pairs of ion images; and an exit questionnaire asking for a feedback both on the task and the survey design. Screenshots of the sections can be found in the GitHub project repository.

The welcome page introduced raters to the survey and its broad aims, described what they are expected to do and how long it can take and provided contact details for the project. The instruction

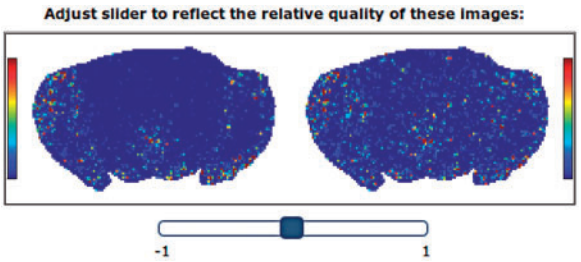

Fig. 2. A screenshot from the online survey showing a pair of ion images. Raters were asked to provide relative quality of these images by moving the slider either to the left or to the right depending on which image they believe is of higher quality

page started with questions about rater background and then detailed how to use the slider to describe the relative quality of two images. Bright 'sticky notes' were used to draw attention to key points such as the individual intensity scaling of each image. To encourage raters to treat the survey as if they were flicking through a dataset, they were instructed to spend at most 10 s per pair.

Image pairs were then presented one-by-one for rating using the following slider mechanic: the two images were displayed side-by-side with a movable slider bar underneath (see Fig. 2), the rater could then indicate which image, if either, of the pair was of higher quality by moving the slider towards the image they considered to be higher quality. For each pair, the slider value set by the user was recorded along with the time spent for providing the rating. Once they had decided on their rating they could click a button to progress to the next page. It was possible to close the survey and return to it later if browser cookies were enabled. An exit questionnaire thanked the rater and gave them opportunity to provide us with feedback on the given task and the survey design. Unless a rater chose to leave their email address during the exit questionnaire the survey was anonymous and no additional information on the rater was recorded.

2.1.2.2 Assignment of image pairs to survey instances. To determine the number of pairs, compile them from images, and design how to show them for rating, we first decided on the number of raters we can involve. Based on our knowledge of experts (subjectively estimated to be around 100) and the response rate of 75% during the pilot study (9 out of 12), we estimated that 75 raters could be recruited to the survey. Then, we decided that each pair has to be rated by three different raters to later assess the agreement [based on typical literature values (Kaufman *et al.*, 2008)]. Then, we decided to show 52 image pairs to a rater (49 different image pairs plus 3 showings of the same pair used only for the consistency check). For a description of the consistency check, please see Section 2.1.4. The number was selected based on our experience in designing surveys and was evaluated in the pilot study (see the next paragraph). This led to $75 \times 49 = 3675$ possible ratings that could be applied to rate $3675/3 = 1225$ unique image pairs, each pair rated 3 times. Finally, the number of ion images we could consider for the survey and for the later gold standard was determined to be 50 as the maximal number of images with ≤ 1225 possible image pairs. Additionally, 2 more images were required for the internal consistency check that led us to selecting 52 ion images as described earlier in Section 2.1.1.

We prepared 75 instances of online surveys with a subset of image pairs to be shown; one survey to be rated by one rater. Each rater was shown 52 pairs; of these 3 were the same pair shown repeatedly for the later consistency check. The same consistency check pair was shown to all raters. All 1225 pairs were randomly allocated into 75 groups, ensuring each pair was shown 3 times and that no survey contained any particular pair more than once. We assigned a default left/right order for the images in each pair. When showing a pair we randomized which image would be presented on the left or right.

2.1.2.3 Online platform for the survey. All the surveys were pre-generated in the required format and deployed onto the LimeSurvey platform (*https://www.limesurvey.org/*), each with a unique URL. A custom redirect script was placed on the recruitment page of the website of the EU FP7 project 3D-MASSOMICS (*http://3d-massomics.eu*) which held a queue of unvisited surveys. When a potential rater followed the link within the invite email they were automatically redirected to the URL of the next unseen survey in the queue. Any surveys that were started but not completed or scored poorly on a set of three consistency control image pairs were manually returned to the queue after 1 week of inactivity.

2.1.2.4 Pilot study. Before the main study, a small pilot survey involving 9 raters (12 were invited, 9 agreed to contribute) was performed to assess whether our task was feasible and to evaluate the survey design. These raters were excluded from being invited to the full survey. In general, the pilot raters were able to comprehend and perform the task. The pilot study made us adjust the scheme of assignment of image pairs to survey instances. Instead of initially planned on-the-fly random sampling of image pairs with replacement, we decided to use pre-generated surveys, since many pairs did not receive the required number of ratings. Some changes to the slider mechanic were also made: the intervals were reduced so the movement was smoother and an explicit instruction that the 0 value could be used was included. We updated the instructions making clear that all images were of coronal sections from mouse brain, each scaled to its maximum value.

2.1.3 Recruitment or raters

Our aim was to recruit raters with substantial expertise in MALDI IMS. We prepared an invitation and distributed it to correspondence emails from journal articles where MALDI IMS was a primary feature and to researchers known to us as being active in this field. Group leaders of IMS research groups were emailed with a request to recruit members of their groups. Towards the end of the survey time, a post was placed on the IMS group on LinkedIn. As an incentive to complete the survey, a small prize was offered to the rater whose performance will be closest to the average of all the expert raters. To be eligible for this, raters must have provided their email address.

2.1.4 Assessment of ratings and compilation of the gold standard

Our aim was to assess the inter-rater agreement and to compile the gold-standard set of image pairs annotated with average ratings.

2.1.4.1 Rating pre-processing. All ratings provided by one rater were normalized by dividing by their standard deviation. This aimed to compensate for inter-raters variations in the scale of provided values, since no training was provided to the raters on the absolute values their sliders should take. The obtained scaled rating was then rounded to the nearest 0.1. Additionally, while the surveys were being collected, the same pair was shown to a rater three times (at position 13, 26 and 38 in the survey) as a consistency check. The standard deviation of the three ratings was calculated and raters

with a standard deviation >0.25 were excluded from the analysis. The corresponding survey instance was returned to the queue to be offered to another potential rater. The slider values were inverted where required to match the default left/right position in each ion image pair.

2.1.4.2 Measuring inter-rater agreement. For measuring the inter-rater agreement, we selected the Krippendorff's alpha that is a standard measure of the agreement fitting our experiment setup (interval ratings, more than two raters). The Krippendorff's alpha is defined as: $\alpha_k = 1 - D_o/D_e$, where D_o and D_e are the measured and expected disagreements between raters, respectively; see Krippendorff (2007) for a thorough explanation of how to perform the calculation. For calculating Krippendorff's alpha, we treated the ratings (slider values) as interval values with 0.1 intervals.

2.1.4.3 Compiling the gold standard. We evaluated whether any particular rater was inconsistent with the cohort that may indicate a low level of expertise, insufficient attention, or problems with a particular instance of the survey. For this, we calculated a set Krippendorff's alpha values in a leave-one-out style with each rater being removed in turn. In a similar way, we evaluated whether ratings for any particular image pair were inconsistent with ratings for all pairs that may indicate unusual images in the pair, complexity of rating this particular pair or problems in presenting this pair in the survey. Finally, a gold-standard subset of image pairs was selected by sequentially removing pairs based on their individual effect on the agreement until the overall agreement value was maximized; each image pair was annotated with the mean of the slider values provided by three raters.

2.2 Potential image quality measures

Once the gold-standard set of pairs of ion images with assigned ratings of relative quality was compiled, we aimed to use it for evaluating candidate measures of quality of ion images which are formulated based on the image analysis principles. Moreover, we aimed to train a composite measure and evaluate whether it can outperform the individual candidate measures.

2.2.1 Candidate measures of quality of ion images

For candidate measures of quality of a real-valued ion image I with n pixels with intensities from [0, 1], we considered those statistics which quantify variation or noise within an image locally as well as globally. In addition, we considered a measure of spatial chaos (SC), which we recently developed (Alexandrov and Bartels, 2013). Each candidate measure was scaled to take values in [0, 1] with low values corresponding to low quality.

2.2.1.1 Local coefficient of variation. For a set of intensity values, the coefficient of variation is defined as the SD (σ) divided by the mean (μ). We considered a local square window of pixels within an image and moved it over the image pixel-by-pixel. For each local window, we calculated local SD and local mean of intensities (σ_{local} and μ_{local}) and taking the ratio to give each pixel its local coefficient of variation:

$$\mathrm{COV}_i = \frac{1}{\sqrt{m-1}}\left\{\frac{\sigma_{local}}{\mu_{local}}\right\}. \quad (1)$$

where i denotes the intensity of ith pixel of an ion image I with i indexing over all n pixels and m is the number of pixels within a window. This measure was also calculated taking the pixel values for the entire image.

2.2.1.2 Local signal-to-noise ratio. For a set of intensity values, the signal-to-noise ratio is defined as the mean divided by SD and is the inverse of the coefficient of variation. Again, we calculated the values of the local standard deviation σ_{local} and local mean μ_{local} within a window:

$$\mathrm{SNR}_i = \frac{1}{\sqrt{m-1}}\left\{\frac{\mu_{local}}{\sigma_{local}}\right\}. \quad (2)$$

This measure was also calculated for the entire image.

2.2.1.3 Local SD. We calculated the local standard deviation within a window σ_{local} for all pixels.

$$\mathrm{STD}_i = 2\{\sigma_{local}\}. \quad (3)$$

This measure was also calculated for the entire image.

2.2.1.4 Square error. We considered another statistic representing the level of noise in an ion image, defined as the mean square error (SE) between an image I and its de-noised version $\bar{I}$. The SE is calculated as:

$$\mathrm{SE}_i = (I_i - \bar{I}_i)^2 \quad (4)$$

The de-noised version of an image was created by applying a Gaussian convolution filter, where we varied the SD of the Gaussian function.

2.2.1.5 Spatial Chaos. In addition to the measures quantifying the level of variation or noise locally or globally, we considered quantifying the amount of SC or noise by using the measure of SC $\mathrm{SC}_{\mathrm{AlexandrovBartels}}$ proposed by Alexandrov and Bartels (2013). We used the default parameters as described in Alexandrov and Bartels (2013) ($\sigma_i = 0.3, \sigma_d = 3, \omega = 10$). In this article, we define the SC measure to be scaled as follows with the aim to have low values for a chaotic image as we assume a noisy image to be of low quality:

$$\mathrm{SC} = 1 - \mathrm{SC}_{\mathrm{AlexandrovBartels}}. \quad (5)$$

2.2.1.6 Image histogram. Histogram measures are frequently used in the analysis of natural scenes (Vogel and Schiele, 2004) as they catalogue the relative abundance of intensities within the image. The greyscale histogram was calculated with bin widths of 0.05. (The intensities of each image I were scaled to [0, 1].)

2.2.1.7 RGB image. As the images were presented to the raters with a colour mapping from their intensity values, and this can affect the judgement of images (Borland and Taylor, 2007), we also explored some colour statistics after mapping the greyscale intensities using the *jet* colourmap in Matlab (Mathworks, USA) with eight-bit colour depth.

2.2.1.8 Luminescence. Image luminescence relates the amount of visual contrast that is present in an image and depends on colour hue and saturation as well as brightness (Webster and Mollon, 1994). In general, more luminescent colours are more noticeable to the human perception. The luminescence of a pixel within an RGB image is defined to be

$$L_i = 0.2126R_i + 0.7152G_i + 0.0722B_i \quad (6)$$

at a pixel i (Akyuz and Reinhard, 2006).

Table 1. Measures calculated per image, where the input can be the raw intensity values or mapped onto a jet colourscale

Input image	Measure	Statistics	Window size
Grey	COV	a	3, 5, 11, 21, 51
	STD	a	3, 5, 11, 21, 51
	SNR	a	3, 5, 11, 21, 51
	SE	a	3, 5, 11, 21, 51
	SC	b	
	histogram	c	
RGB	luminescence	a	
	histogram (per channel)	c	

Sets of summary statistics calculated were:a, mean, median, maximum absolute deviation (mad), maximum (max), minimum (min), sum; b, mean; c, skew, kurtosis, entropy, maximum value. Each measure was applied to the whole image, and if indicated in the column 'Window size', locally to moving square windows of specified size in pixels.

2.2.1.9 RGB histograms. The frequency of intensities was calculated for each Red, Green and Blue channel independently.

2.2.1.10 Window sizes and summary statistics. The measures COV, STD, SNR and SE are non-local in the sense that at some point during their computation they take information from a window of surrounding pixels. To understand whether the degree of non-locality relates to human perception of the images we varied the window width between 3 and 51 pixels to cover all length scales. As we wished for each combination of measure and window to return a single assessment value we calculated a set of image statistics for each measure as summarized in Table 1. This led to considering 143 features calculated per image.

2.2.2 Evaluating candidate measures of quality of ion images

We evaluated how well the candidate measures reproduced the expert ratings. For each measure we defined a differential for every pair of images: the measure was calculated for the left and right images independently and then the value for the left was subtracted from the right. This was calculated for every pair in the survey and two metrics were then calculated between the differential measures and the expert ratings. First, we calculated the sample Pearson correlation coefficient to examine scale invariant trends on an absolute scale; second, we compared whether the sign (positive or negative) of the measures matched that of the expert raters:

$$s = \frac{1}{n}\sum_{i=1}^{n} P(x_i, y_i),\ P(x,y) = \begin{cases} 1 \text{ if } & x \le 0 \,\&\, y \le 0 \\ 1 \text{ if } & x > 0 \,\&\, y > 0 \\ 0 & \text{otherwise} \end{cases}. \quad (7)$$

2.2.3 Composite measure of quality of ion images

We investigated whether a composite of the candidate measures could outperform each of them. For this, we considered a linear combination of the proposed candidate measures and their multipliers were optimized by minimizing the least squared difference between the composite measure and the gold-standard annotations with the lasso regularization. The lasso regularization was employed to select the most relevant features out of groups of similar features. The solver *lasso* in Matlab (Mathworks, USA) was used with the 10-fold cross validation.

3 Results and discussion

3.1 Gold-standard obtained through expert rating

3.1.1 Selected ion images

Given the pool of images obtained as described in Section 2.1.1, two images were manually selected for an internal consistency check and 50 for the online survey itself. In total, 1225 image pairs were generated for the survey. Our knowledge of MALDI-IMS suggested that images from the same *mz* in datasets with different sample-handling were likely to have different relative noise characteristics. Comparison of the mean spectra between the datasets (data not shown) revealed reasonable heterogeneity in peak intensities indicating that the expected variation was achieved. The mean spectra with 49 peaks used for the survey images and the images itself can be found in the project GitHub repository.

3.1.2 Recruited raters

Based on the response times for the pilot study, we allocated one month for recruitment, but it took longer than expected to involve enough raters. We extended the recruitment time to three months with two follow-up email requests and a post on the LinkedIn IMS group. In total, 80 experts completed the survey.

All recruited experts indicated some experience with MALDI IMS data, with total experience over 260 years. Nine raters reported more than ten years experience each. The distribution of experience and work background is shown graphically in Figure 3a. The majority of raters (45%) had a background of using IMS primarily for investigating biological or biochemical problems, 24% for data acquisition and the remainder encountered IMS data during technology or algorithm development.

3.1.3 Interaction of raters with the online survey

The feedback from both the pilot and the final survey raters was that the slider mechanic was easy to understand and to use (see Fig. 3b). The exit questionnaire revealed that four raters had difficulty in using the slider, two of these thought the granularity of the slider was too fine and the other two found it difficult to set the value to exactly zero.

It was agreed by 95% of raters that completed the survey that the task determining the quality of images from IMS was important: agreed or strongly agreed with the statement 'Determination of image quality is an important part of a MALDI imaging experiment'; see Figure 3b. However, the statement 'It was easy to decide on the relative quality of a pair of images' proved divisive amongst users with the majority of responses neither agreeing or disagreeing with the statement. In contrast to the survey mechanisms, which were generally agreed to be easy to understand, the task of actually rating a pair of images was not found to be straightforward. This reinforces the importance of finding a general measure that can replace subjective judgements and present route to a standardized image quality score for reporting purposes.

We recorded how long each rater spent on rating each pair, see Figure 4d. The obtained learning curve is in line with our expectations: the raters were spending more time per rating when starting the survey with less time per rating being needed as the survey progressed. If this average time was extrapolated to a typical MALDI-TOF dataset with 10 000 channels then manually going through each to determine the quality would take over 20 h and high-resolution IMS would take many days. This quantifies the amount of human time needed for the analysis of IMS data, which has been noted as a rate limiting step (Alexandrov, 2012; Pacholski and Winograd, 2010; Palmer *et al.*, 2013).

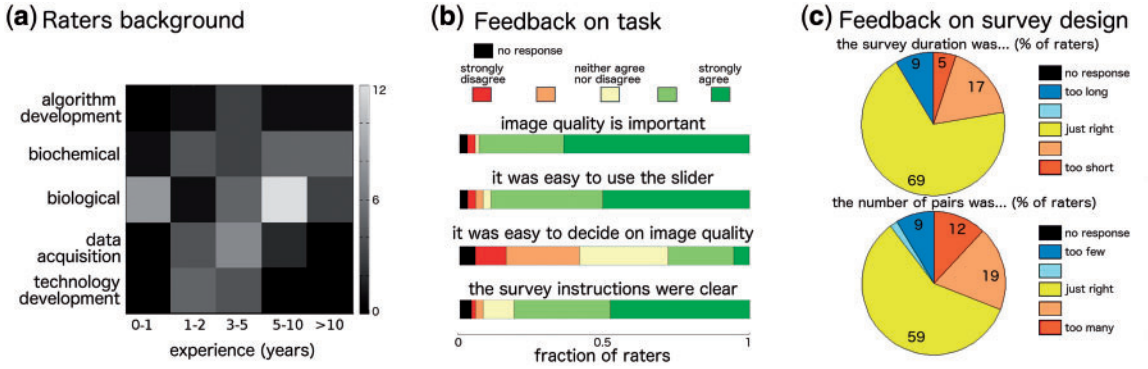

Fig. 3. Overview of the raters who completed the survey and their feedback on the task and survey. **(a)** Raters' experience and background show that a diverse range of experts was recruited. **(b)** The raters described the objectives of the survey as important, the mechanics and the layout of the survey clear and easy to use but the task of determining image quality was found to be difficult. **(c)** Raters' feedback on the survey duration and number of pairs showed that the survey was comfortable for the participants

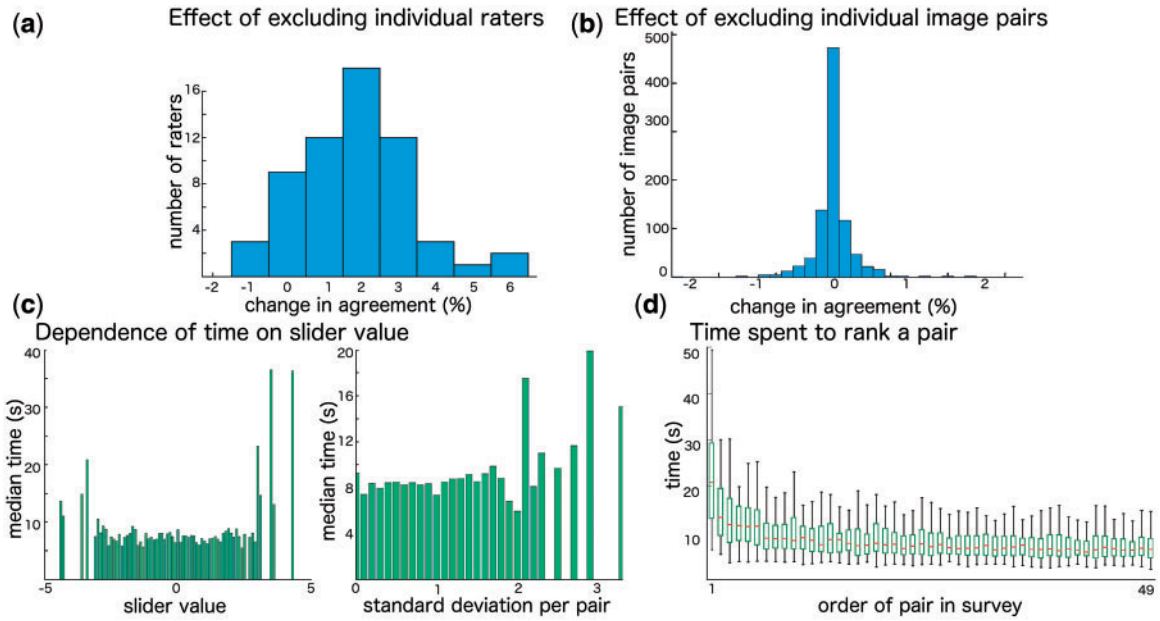

Fig. 4. Assessment of the raters and the ratings they provided. **(a)** A histogram of the change of the inter-rater agreement when removing one rater in turn. **(b)** A histogram of the change of the inter-rater agreement when removing one image pair in turn. **(c)** The median time spent per pair against average slider value and SD per pair. **(d)** Box-whisker plot of tracked time shows an expected learning curve with less time per rating being needed as survey progresses (plot shows 25/50/75 quantiles with whiskers covering 99.3% of data)

3.1.4 Assessment of ratings and obtained gold standard

3.1.4.1 Consistency check. Out of 80 recruited experts, 20 did not pass the consistency check (see Section 2.1.4). The ratings produced by these experts were excluded from further consideration. As a result, we obtained ratings from 60 raters which produced 634 image pairs with three ratings.

3.1.4.2 Inter-rater agreement. The Krippendorff's alpha α_k value for the 60 raters was 0.5. This was lower than expected from the pilot study where a value of 0.65 was achieved and below the usually accepted threshold of 0.6 (Krippendorff, 2012).

It was important for us to understand why the inter-rater agreement was low in order to judge whether the ratings obtained would still be useful. We considered three hypotheses which could lead to low agreement between raters: (i) some raters performed badly compared with the cohort, (ii) there were pairs within the survey that were too difficult to rate, (iii) the whole task was not achievable. We were confident that the task was feasible as our pilot study achieved a reasonable α_k value and so we assessed the first two hypotheses in more detail.

We calculated α_k with each rater excluded in turn to see if individual raters performed differently to the rest of the cohort. In general, removing specific raters had a small impact (<5%) on the value of α_k (see Fig. 4a), but removing two raters caused an improvement. Removing both of these raters from the cohort increased the value of α_k to 0.55, still lower than could be called a sufficient agreement. We compared the median time each expert spent on the ratings against the change in α_k when removing this expert, but no correlation was visible (data not shown). We then scored raters according to their impact on the overall agreement and sequentially

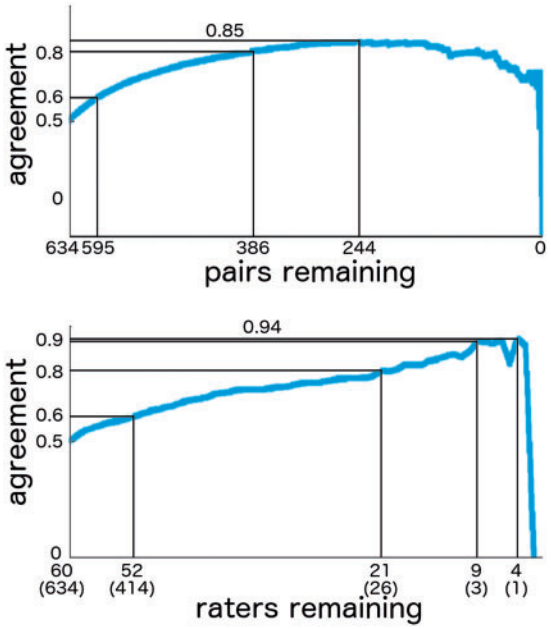

Fig. 5. Agreement (Krippendorff's alpha α_k) calculated after sequentially removing the worst performing image pairs (top) and raters (bottom) then recalculating the agreement on the remaining subset (i.e. removing from left to right in Fig. 4a and b)

removed the worst performing raters. As shown in Figure 5 (bottom), removing raters also resulted in a gradual improvement of the agreement. The agreement of 0.6 and 0.8 was reached with 414 and 26 image pairs left, correspondingly.

We then calculated α_k with each image pair excluded in turn. As the number of image pairs is very large compared with the number of raters it would not be expected that removing individual pairs would have a substantial impact on the overall agreement levels. However, for a few pairs it led to improving the value of α_k by over 1%; see Figure 4b. Plotting agreement calculated with removed pairs against the median time did not reveal any correlation, suggesting that those pairs with low agreement were not inspected for any more or less time than other pairs. We then scored image pairs according to their impact on the overall agreement. Sequentially removing the worst performing pairs steadily improved the agreement value until a maximum was reached after which a plateau was seen; see Figure 5 (top). The agreement of 0.6 and 0.8 was reached with 595 and 386 image pairs left, correspondingly. The maximum agreement of 0.85 was reached with 244 image pairs and taking responses from 60 raters.

These observations suggest that there were both problematic raters and image pairs difficult to rate. Another conclusion is that it proved to be possible to isolate a set of high-agreement image pairs.

3.1.4.3 Gold standard. The key feature of a gold-standard set is that it received a consistent judgement by all the experts who rated it. Based on the observations presented in the previous paragraph, we decided to exclude low-agreement image pairs from the gold standard. Alternatively, raters, which performed worse with respect to the overall agreement, could be excluded, but it would result in a much smaller gold standard, because excluding one rater implies excluding all image pairs that they have rated.

Given the agreement results, we generated three gold-standard datasets based on three different agreement thresholds (see Fig. 5). The first dataset consisting of 595 image pairs includes pairs rated with the usually accepted agreement of 0.6. The second agreement threshold is defined as a standard deviation of normalized slider values (386 image pairs). The third dataset consists of 244 image pairs that produced the maximal agreement of $\alpha_k = 0.85$.

Future work includes a more detailed study of features of the high- and low-agreement image pairs that will allow us to draw conclusions about the reasons for the raters to disagree on certain subsets of images.

The gold-standard datasets can be found in the GitHub repository as a freely available resource to provide the community with an annotated dataset that we hope will spur further developments in this area.

3.2 Reproducing human judgements with proposed measures

Human assessment of image quality is a time consuming task, particularly if consensus between several experts is required. In our study, a total of 12.5 person-hours were required to get 3675 ratings. We considered several candidate measures which took into account both local variation and the whole image and evaluated their ability to reproduce human judgements about quality of MALDI ion images. The measures looked at heterogeneity (*COV*, *STD*), relative intensity (*SNR*), global noise (*SE*), gradient patterns (*SC*), scene statistics (histogram) and image colour (luminescence, RGB histograms). The *COV*, *STD* and *SNR* operate locally whilst the *SE*, *SC* and histogram operate on an entire image to return a number quantifying the global level of noise or the degree of structure present in the image; (see Alexandrov and Bartels, 2013) for a discussion on the structure and chaos in MALDI ion images.

3.2.1 Performance of candidate measures

The values of differential measures were compared with the expert ratings on each of the three gold-standard sets defined by taking the agreement cut-off values of $\alpha = 0.5$, $\alpha = 0.6$ and $\alpha = 0.85$; see Figure 5a. Two statistics were used to evaluate the similarity of each measure to the human ratings of MALDI ion image quality. First, we calculated the sample Pearson correlation coefficient between the differential measure values and the ratings in the gold standard. The Pearson correlation is a standardized measure of linear relationship and so is easy-interpretable and comparable. Second, we compared the sign of the differential measure with the sign of the rating. This provides a simple classification of which image is of higher or lower quality, independent on the magnitude of the difference. The top 10 performing statistics are summarized in Table 2; the full table can be found in the GitHub repository.

This table reveals that two measures (*mad*STD and SC) showed a strong linear relationship with the expert assessments for all three gold-standard datasets, and that the linearity increased with the rater agreement. This may indicate that local variation and the spatial structure are the characteristics that the expert raters use to judge when assessing the relative quality of images. The sign match accuracy shows a similar trend, increasing from >0.7 to >0.9; all values are >0.5 that would be produced by simple guessing. It is noteworthy that among the top 10 performing measures only the one measure ($medianSTD_5$) operates on a small moving window (5×5). All other measures operate either on a full image (*SC*) or on a large moving windows (at least 11×11 in size); four measures operate on 51×51 moving windows. This suggests that the expert raters were taking larger image areas into account, rather than very small details. The majority of the top performing measures were

Table 2. Evaluation of the candidate image-based measures on the three gold-standard datasets

Measure	$\alpha = 0.5$		$\alpha = 0.6$		$\alpha = 0.85$	
	Corr*	Sign	Corr*	Sign	Corr*	Sign
madSTD_{11}	0.74	0.76	0.76	0.77	0.93	0.93
madSTD_{21}	0.71	0.77	0.74	0.78	0.93	0.93
SC	0.58	0.74	0.59	0.75	0.72	0.85
madSTD_{51}	0.54	0.70	0.56	0.71	0.67	0.82
madSNR_{51}	0.48	0.67	0.49	0.67	0.61	0.76
minSTD_{21}	0.48	0.67	0.49	0.68	0.61	0.75
madCOV_{51}	0.44	0.62	0.45	0.62	0.57	0.69
medianSTD_{5}	0.40	0.63	0.41	0.63	0.54	0.72
medianSTD_{11}	0.40	0.63	0.41	0.63	0.53	0.71
maxCOV_{51}	0.40	0.62	0.41	0.61	0.53	0.71

Out of 143 considered measures, 10 with the highest correlation are shown, sorted by their Pearson correlation (Corr) between the differential measures and the human ratings. Sign stands for the sign matching statistic as defined in Equation (7). *All P-values < 0.001.

based on the maximum absolute deviation, a robust measure of the spread of values, or the median indicating that raters may have preferred images where there was consistency in the local noise across an image.

3.2.2 Performance of a composite measure trained on the gold standard

We investigated whether a composite measure can better reproduce the human judgements. We considered a linear combination of measures and optimized the linear multipliers as described in Section 2.2.3 noting that the lasso approach is expected to eliminate duplicitous features. The gold standard with $\alpha_k = 0.85$ was used. The optimum combination (in a least squares sense) produced a set of 18 measures that was largely composed of *COV*, *SNR* and *STD*. This linear combination produced a linear correlation (Corr) of 0.94 and a sign match accuracy (Sign) of 0.98. This represents a very modest improvement over the top performing individual algorithms (achieved Corr of 0.93 and Sign of 0.93). There was a substantial variation in the measures included in the top scoring sets for different initialization values that is indicative of a complex search space with multiple local minima.

Based on the results of the evaluation, we recommend for quantifying the quality of MALDI ion images to use either the *mad*STD_{11} or *SC* measure as these consistently outperformed other candidate image-based measures, or consider using the composite measure with optimized linear multipliers. Note that the composite measure may not generalize to all sample types, mass spectrometry systems used and experimental settings as it was trained on the set of images chosen for this survey, whereas no parameters optimization was performed for the *SC* measure.

4 Conclusion

In recent years, IMS emerged as an indispensable tool of molecular imaging. With the increasing rate of IMS data collected, the number of laboratories performing the experiments, and the rising complexity of protocols, there is a growing need for an objective and quantitative measure of quality of IMS data that would account for specific properties of IMS data, in particular for imaging information inherently present in IMS data. Having such a measure would open up avenues for optimization and monitoring of data acquisition. It could be included in-line with automated data acquisition to monitor instrument performance and provide warning of low-quality data collection so that appropriate steps can be taken to restore the instrument to full capacity. The images included in the survey were on the lower end of the quality of data that can be produced and the next stages would be to assess how raters interpret the relative quality of high-resolution data.

We proposed new measures based on the image analysis principles, created a gold-standard set of ion images annotated with relative ratings of quality by a crowd of experts, and illustrated how the gold-standard set can be used to derive a composite measure of quality of ion images.

Unexpectedly for us, involving a crowd of experts and creating a gold-standard set of ion images turned out to be a significant challenge that can explain why this approach was not yet applied in the field of analytical chemistry. However, despite the challenges faced we continue to believe that this approach provides the best way to assessment of quality of IMS data through formalization of subjective knowledge possessed by experts. Our approach demonstrated that despite the lack of formal definition of quality of IMS data, this knowledge is indeed possessed by experts. By publishing the gold-standard set of relatively rated pairs of ion images together with the source code for the data analysis, we invite the community to test their own algorithms and compare the performance to our results.

Acknowledgements

We thank all experts who contributed by rating image pairs. We thank the reviewers for their detailed comments and suggestions that helped us substantially improve the article.

Funding

This work was supported by European Union 7th Framework Program grant (305259).

Conflict of Interest: none declared.

References

Abdel-Rehim,M. (2004) New trend in sample preparation: on-line microextraction in packed syringe for liquid and gas chromatography applications I. Determination of local anaesthetics in human plasma samples using gas chromatography-mass spectrometry. *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.*, **801**, 317–321.

Akyüz,A.O. and Reinhard, E. (2006) Color appearance in high-dynamic-range imaging. *J. Electron. Imaging*, **15**, 033001.

Alexandrov,T. (2012) MALDI imaging mass spectrometry: statistical data analysis and current computational challenges. *BMC Bioinformatics*, **13**(Suppl 16), 1–13.

Alexandrov,T. and Bartels,A. (2013) Testing for presence of known and unknown molecules in imaging mass spectrometry. *Bioinformatics*, **29**, 2335–2342.

Balluff,B. *et al.* (2015) De novo discovery of phenotypic intratumour heterogeneity using imaging mass spectrometry. *J. Pathol.*, **235**, 3–13.

Bern,M. *et al.* (2004) Automatic quality assessment of peptide tandem mass spectra. *Bioinformatics*, **20**, 49–54.

Berner,E.S. (2003) Diagnostic decision support systems: how to determine the gold standard? *J. Am. Med. Informatics Assoc.*, **10**, 608–610.

Borland,D. *et al.* (2007) Rainbow color map (Still) considered harmful. *IEEE Comput. Graph. Appl.*, **27**, 14–17.

Bradley,J.C. *et al.* (2009) The spectral game: leveraging open data and crowdsourcing for education. *J. Cheminform.*, **1**, 9.

Carletta,J. (1996) Assessing agreement on classification tasks: the kappa statistic. *Comput. Linguist.*, **22**, 249–254.

Chaurand,P. (2012) Imaging mass spectrometry of thin tissue sections: a decade of collective efforts. *J. Proteomics*, **75**, 4883–4892.

Coombes,K.R. *et al*. (2005) Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. *Proteomics*, **5**, 4107–4117.

Deininger,S.-O.O. *et al*. (2011) Normalization in MALDI-TOF imaging datasets of proteins: practical considerations. *Anal. Bioanal. Chem.*, **401**, 1–15.

Dresen,S. *et al*. (2010) Detection and identification of 700 drugs by multitarget screening with a 3200 Q TRAP LC-MS/MS system and library searching. *Anal. Bioanal. Chem.*, **396**, 2425–2434.

Du,L. *et al*. (2014) Crowdsourcing natural products discovery to access uncharted dimensions of fungal metabolite diversity. *Angew. Chemie Int. Ed.*, **53**, 804–809.

Gwet,K.L. (2012) *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Multiple Raters*. Advanced Analytics Press, Gaithersburg.

Karlsson,O. *et al*. (2014) Quality measures of imaging mass spectrometry aids in revealing long-term striatal protein changes induced by neonatal exposure to the cyanobacterial toxin β-N-methylamino-L-alanine (BMAA). *Mol. Cell. Proteomics*, **13**, 93–104.

Kaufman,J.C. *et al*. (2008) A comparison of expert and nonexpert raters using the consensual assessment technique. *Creat. Res. J.*, **20**, 171–178.

Krippendorff,K. (2007) Computing Krippendorffs alpha reliability. *Communication*, **43**, 1–9.

Krippendorff,K. (2012) *Content Analysis: An Introduction to Its Methodology*. Sage, Thousands Oaks.

Ma,B. *et al*. (2003) PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.*, **17**, 2337–2342.

McDonnell,L.A. *et al*. (2014) MSiMass list: a public database of identifications for protein MALDI MS imaging. *J. Proteome Res.*, **13**, 1138–1142.

Mutton,I. *et al*. (2011) The design and use of a simple System Suitability Test Mix for generic reverse phase high performance liquid chromatography-mass spectrometry systems and the implications for automated system monitoring using global software tracking. *J. Chromatogr. A*, **1218**, 3711–3717.

Nesvizhskii,A.I. *et al*. (2006) Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. *Mol. Cell. Proteomics*, **5**, 652–670.

Oetjen,J. *et al*. (2013) MRI-compatible pipeline for three-dimensional MALDI imaging mass spectrometry using PAXgene fixation. *J. Proteomics*, **90**, 52–60.

Pacholski,M.L. and Winograd,N. (2010) Mass spectrometry imaging. *Methods Mol. Biol.*, **656**, 99–111.

Palmer,A.D. *et al*. (2013) Randomized approximation methods for the efficient compression and analysis of hyperspectral data. *Anal. Chem.*, **85**, 5078–5086.

Römpp,A. *et al*. (2014) A public repository for mass spectrometry imaging data. *Anal. Bioanal. Chem.*, **407**, 2027–2033.

Schuerenberg,M. and Deininger,S.O. (2010) Chapter 7: matrix application with ImagePrep. In: Setou, M. (ed.) *Imaging Mass Spectrometry: Protocols for Mass Microscopy*. Springer, Tokyo, pp. 87–91.

Spengler,B. (2015) Mass spectrometry imaging of biomolecular information. *Anal. Chem.*, **87**, 64–82.

Trede,D. *et al*. (2012) Exploring three-dimensional matrix-assisted laser desorption/ionization imaging mass spectrometry data: three-dimensional spatial segmentation of mouse kidney. *Anal. Chem.*, **84**, 6079–6087.

Vogel,J. *et al*. (2004) Natural scene retrieval based on a semantic modeling step. *Image Video Retr. Lect. Notes Comput. Sci.*, **3115**, 207–215.

Watrous,J.D. *et al*. (2011) The evolving field of imaging mass spectrometry and its impact on future biological research. *J. Mass Spectrom.*, **46**, 209–222.

Webster,M.A.M. *et al*. (1994) The influence of contrast adaptation appearance on color appearance. *Vision Res.*, **34**, 1993–2020.