International Journal of Mass Spectrometry 437 (2019) 49–57

Contents lists available at ScienceDirect

International Journal of Mass Spectrometry

journal homepage: www.elsevier.com/locate/ijms

# Statistical detection of differentially abundant ions in mass spectrometry-based imaging experiments with complex designs

Kylie A. Bemis[a,1], Dan Guo[a,1], April J. Harry[b,c,1], Mathew Thomas[d], Ingela Lanekoff[f], Mary P. Stenzel-Poore[e], Susan L. Stevens[e], Julia Laskin[g], Olga Vitek[a,*]

[a] College of Computer and Information Science, Northeastern University, Boston, MA, USA
[b] Department of Statistics, Purdue University, West Lafayette, IN, USA
[c] College of Science, Northeastern University, Boston, MA, USA
[d] Pacific Northwest National Laboratory, Richland, WA, USA
[e] Department of Molecular Microbiology & Immunology, Oregon Health & Science University, Portland, OR, USA
[f] Department of Chemistry – BMC, Uppsala University, Uppsala, Sweden
[g] Department of Chemistry, Purdue University, West Lafayette, IN, USA

ARTICLE INFO

*Article history:*
Received 11 August 2017
Received in revised form 24 April 2018
Accepted 23 July 2018
Available online 3 August 2018

*Keywords:*
Mass spectrometry imaging
Nano-DESI MSI
DESI MSI
Experimental design
Statistical analysis
Spatial statistics

ABSTRACT

Mass Spectrometry Imaging (MSI) characterizes changes in chemical composition between regions of biological samples such as tissues. One goal of statistical analysis of MSI experiments is class comparison, i.e. determining analytes that change in abundance between conditions more systematically than as expected by random variation. To reach accurate and reproducible conclusions, statistical analysis must appropriately reflect the initial research question, the design of the MSI experiment, and all the associated sources of variation. This manuscript highlights the importance of following these general statistical principles. Using the example of two case studies with complex experimental designs, and with different strategies of data acquisition, we demonstrate the extent to which choices made at key points of this workflow impact the results, and provide suggestions for appropriate design and analysis of MSI experiments that aim at detecting differentially abundant analytes.



## 1. Introduction

Mass spectrometry-based imaging (MSI) characterizes and visualizes the chemical makeup of biological tissues [1,2]. A key advantage of MSI is its spatial registration, which provides unique information for a variety of applications, such as histological examination of specimens [3], delineating tumor margins in surgical settings [4], characterizing sample heterogeneity down to cellular resolution [5], or drug mapping [6–8]. Images from serial tissues can be used to characterize the chemical heterogeneity of three-dimensional volumes [9], or examined against images of the same tissue obtained by other modalities, e.g. with MRI [10,11]. Such applications distinguish MSI from other mass spectrometry-based experiments.

MSI is performed in a variety of workflows. For example, Matrix-Assisted Laser Desorption Ionization (MALDI) [12] requires the application of a matrix over the tissue, while Desorption Electrospray Ionization (DESI) [13] and nanospray Desorption Ionization (nano-DESI) [14] require less sample preparation. With all the workflows, mass spectra are acquired at gridded locations across the tissue. The spatial distribution of the individual analytes, represented by their mass to charge ($m/z$) ratios is then visualized as molecular ion images, and analyzed using computational and statistical methods [15,16].

Different scientific goals of MSI experiments can be translated into different statistical goals. The typical statistical goals are *class discovery* (i.e., image segmentation, e.g. for characterizing tissue morphology) and *class prediction* (i.e., image classification, e.g. for discovery of biomarkers of disease). A less frequent, but nevertheless equally important goal is *class comparison*. Class comparison considers regions or conditions pre-defined by an external reference, such as a pathological examination, located in a same tissue or across different tissues. Class comparison detects *differentially abundant* ions, i.e. ions that change in average abundance in the individuals from the underlying populations, more systematically than as expected by random chance. Since class comparison considers averages across individuals, it is not appropriate for discovery

* Corresponding author.
*E-mail address:* o.vitek@neu.edu (O. Vitek).
[1] Equal contribution.

https://doi.org/10.1016/j.ijms.2018.07.006
1387-3806/

of biomarkers of disease. Differentially abundant analytes are not necessarily predictive of an individual subject's status of the disease.

Class comparison of biological samples must take into account the variation from multiple sources (which may be absent in non-biological samples such as fingerprints or documents [17,18]). First, there is a natural variation in analyte abundance across distinct biological individuals. Second, variation is present within a same biological tissue, reflecting the heterogeneity of its chemical composition. Finally, experimental and technological artifacts also contribute to the overall variation.

Statistical design and analysis of experiments offers a principled way of accounting for all of these sources when determining differential abundance [19]. *Statistical experimental design* specifies the conditions of interest, as well as the type and the number of biological replicates. MSI experiments can support complex experimental designs. Conditions and replicates can occur in a variety of combinations, including comparisons of conditions within or between tissues, in a single or multiple biological samples subjects. The goal of the experimental design is to optimize the selection of tissues and replicates, and the protocol of data acquisition, to maximize our ability to detect differentially abundant ions under the constraints of sample availability and cost.

*Statistical analysis of experiments* specifies the protocol of data processing, including details of steps such as baseline subtraction, peak detection and alignment, and normalization. It further specifies a statistical model that describes the data, and a procedure for deriving model-based conclusions. To reach accurate and reproducible conclusions, statistical analysis must appropriately reflect the research question, the experimental design, and all the associated sources of variation.

Unfortunately, statistical design and analysis of MSI experiments has so far received limited attention. This manuscript contributes to this area of research. Focusing on the problem of detecting differentially abundant analyses between pre-defined conditions, and using the example of two case studies with complex experimental designs, we highlight the importance of choices made at key points of the workflow, demonstrate the extent to which these choices impact our ability to detect differentially abundant ions, and provide suggestions for appropriate design and analysis of mass spectrometry-based imaging experiments.

## 2. Background

### 2.1. Statistical experimental design

Statistical experimental design defines conditions or treatments of interest (such as tissue region types, disease status, or treatment group) to be compared, and the type and number of biological replicates. For example, a *group comparison design* represents each condition by tissues from different biological individuals. It can be extended by *subsampling*, i.e. a design that includes multiple tissue sections from a same individual in a condition. Examples of MSI experiments with group comparison design are in [20,21]. Alternatively, a *paired design* represents each condition by regions of a same tissue, or multiple tissues from a same biological individual, over multiple biological individuals. These designs are most effective, as they use each biological individual as its own control, while fully characterizing the natural biological variation in the underlying subject population. Examples of MSI experiments with a paired design are in [22,23].

A special case of paired design is an *unreplicated experiment*, i.e. experiment comparing conditions within a single tissue. Such design is undesirable, as it fails to characterize the biological variation, and reduces the scope of conclusions in a way that is only valid for one tissue. Therefore, it leads to *overfitting*, i.e., mistaking subject-specific artifacts for differences in the entire populations, and to irreproducible results.

A second important aspect of statistical experimental design is the allocation of the tissues to all the steps of sample handling and data acquisition. In particular, randomization of biological replicates prevents biases due to technological artifacts from known and unknown sources [15,19]. Randomization is impossible in unreplicated experiments, and this further undermines their value.

This manuscript contrasts group comparisons and paired design in experiments with tissues from single and multiple biological individuals.

### 2.2. Data processing

Data processing is a protocol of data transformation, normalization and reduction [24], adapted to each process of data acquisition. For example, spectra in MALDI-TOF experiments may contain matrix-specific artifacts. Baseline reduction can help alleviate these artifacts. In nano-DESI experiments, the data are commonly acquired as a collection of line scans, where the time to acquire a mass spectrum varies. Therefore, the analyzed locations are unevenly spaced within a line, and their number varies between lines and tissues. In these experiments, data processing uses acquisition time and average raster speed to infer the coordinates of the analyzed locations, and spectral resampling and interpolation to generate raster images [25].

Spectra from MSI experiments are often reduced to a set of peaks, either by a peak-picking algorithm [26] or by selecting specific compounds of interest. The peaks are then aligned across spectra and possibly filtered, to eliminate peaks with low signal, or appearing in a small proportion of the spectra [24]. As an alternative to peak-picking, the mass spectra may be reduced by binning, summing over the intensities in the bins. Bins may be chosen with a fixed width, or with width that increases with the ratio of mass to charge to account for shifts at large mass values [27].

Tissue composition, instrument variation, or even sample preparation can cause ion suppression, and incomparable mass spectra across locations and tissues. Normalization of the mass spectra to a same scale aims to alleviate these artifacts. The most common normalization equalizes the total ion current (TIC) across locations, to account for global intensity variation [28,29]. While the TIC normalization is adequate for most homogeneous samples with little variation, it can be hampered in heterogeneous tissues with biological variation in high-intensity ions [30]. An alternative is the normalization to internal standards, i.e. molecules known to be homogeneously distributed across tissue locations [31,32]. This normalization aims to correct the artifacts of data acquisition in a way that is not affected by high-intensity ions. Many other normalizations exist, such as those based on median, root mean square, and noise level [28], variance stabilizing normalization [30], sliding window normalization [33], and others [29].

An optimal choice of normalization is study-specific, and reflects its biological and technological characteristics. This manuscript does not aim at suggesting optimal data processing. Instead, we illustrate the extent to which the choice of normalization affects the results of downstream statistical analyses.

### 2.3. Statistical analysis of MSI experiments

According to the general principles of statistical design and analysis of experiments, statistical analysis must match the research goal, the experimental design, and the sources of variation that affect the analytes. We detail these points in the context of MSI experiments.

**Statistical analysis must match the research goal.** The research goal of class comparison is to detect systematic changes in chemical composition between conditions. This translates into the statistical goal of testing the hypothesis $H_0$ of no difference in analyte abundance between the conditions, on average in the underlying biological populations and tissues, against the alternative $H_a$ that a difference exists. This statistical goal is achieved with *supervised analysis*, where locations on the tissues in the experiment are annotated with their conditions. It also requires a statistical model, as we discuss below. The hypotheses are tested separately for each analyte, and are sometimes referred to as "univariate comparisons" [15]. The results are characterized in probabilistic terms, such as statistical power and False Discovery Rate among the analytes as in [34].

Alternative research goals of MSI experiments translate into different statistical goals, and require different statistical analyses. For example, finding regions with homogeneous chemical composition translates into the statistical goal of class discovery. It is achieved with *unsupervised analysis*, e.g., Principal Component Analysis (PCA) [35]. Since unsupervised analysis takes as input unlabeled tissues, it is inappropriate for detecting changes between known conditions.

As another example, predicting the condition of each individual location on a tissue, or of the entire tissue, translates into the statistical goal of class prediction. This is the only statistical goal appropriate for discovery of biomarkers of a disease. Similarly to class comparison, this statistical goal is achieved with *supervised analysis*. Unlike the class comparison, it requires machine learning algorithms such as Partial Least Squares Discriminant Analysis (PLS-DA) [36] to make predictions for each individual biological subject and tissue, while simultaneously using all or a subset of the analytes. Methods such as PLS-DA are inappropriate for studies of differential abundance, because statistical properties of predictive analytes differ from those of differentially abundant analytes.

**Statistical analysis must match the experimental design.** Statistical experimental design defines the regions of the tissues that we would like to compare. The regions can be defined using images acquired with alternative modalities, such as MRI [10,11], optical microsopy [37], or using marker analytes, acquired as part of the MSI experiment. It is important to emphasize that these markers must be defined in advance (as opposed to be used twice - first to determine the regions and then to test for differential abundance between these regions) to avoid overfitting and unduly optimistic conclusions of hypothesis testing.

Tissues selected for MSI experiments can form a group comparison or a paired design, with or without biological replication. However, the advantages of complex designs, such as paired design or design with biological replicates, are lost if they are not accordingly analyzed. For example, a statistical analysis ignoring the paired structure of a design misses the opportunity to use each biological subject as its own control, thus limiting the effect of subject-to-subject variation which is not of interest in the experiment [38].

**Statistical analysis must match the existing sources of variation.** Statistical analysis requires a statistical model, i.e. an abstraction describing the systematic variation between the conditions, and nuisance variation from sources beyond the experimental control. Nuisance variation includes the biological between-subject and within-subject variation, and the measurement error.

Linear mixed effects model is a general and flexible class of such abstractions, applicable to complex designs. A special case is Analysis of Variance (ANOVA), which describes the variation among multiple groups. ANOVA-based statistical analysis of two groups of independent biological replicates is equivalent to a two-sample *t*-test with pooled variance. Models for more complex designs may include, e.g., random effects that distinguish biological variation and measurement error [39,40].

A key assumption of the linear mixed effects models above is that the replicates in the experiment are statistically independent. This is typically true with respect to the biological replicates, which are distinct tissue donors. However, the assumption does not hold with respect to locations in a biological tissue. Chemical composition of biological tissues is prone to *spatial autocorrelation*, i.e., situation where proximate locations in a tissue are more likely to have similar chemical composition than distant locations in the same tissue [41]. One solution to this challenge is to only model one value from each tissue, such as the average abundance of the analyte within each condition and each biological replicate [15,42]. Although this approach is compatible with linear mixed models, it ignores the heterogeneity of the tissue.

An alternative approach is to extend the linear mixed model in a way that explicitly describes spatial autocorrelation. Cassese et al. [41] proposed such a modeling and testing framework, however it is limited to a single tissue, and is not directly applicable to complex experimental designs. Another extension, called Hierarchical Bayesian Spatial Model, combines the flexibility of describing the experimental designs, the independence of biological replicates, and the within-tissue spatial autocorrelation [43]. We illustrate the impact of the modeling choices on the statistical analysis using two case studies below.

## 3. Methods

### 3.1. RCC: DESI-MSI of human renal cell carcinoma

The study was first reported in [22], and discussed in [44]. One of the goals of the study was to detect differentially abundant analytes between cancerous and healthy tissues.

**Statistical experimental design.** The experiment was conducted in a paired design. Pairs of tissues exhibiting renal cell carcinoma (RCC) and adjacent normal tissue were collected from eight human volunteers (Fig. 1). The tissues were subjected to serial hematoxylin and eosin (H&E) staining. The pathology examination of the stained tissues was unable to define homogeneous sub-regions of the tissues with respect to the disease at sufficiently high resolution. Therefore the entire tissue sections were labeled as either "cancer" or "normal".

**Data acquisition.** The tissues were analyzed with DESI MSI in negative mode on a Thermo LTQ Orbitrap instrument. The data were recorded in RAW file format, converted to the Analyze 7.5 format, and imported into the open-source software *Cardinal* v1.10.0 [45]. The resulting dataset is available in the R package *CardinalWorkflows* on Bioconductor.

**Data processing.** All the processing was done using *Cardinal*. Since no high-intensity features with high biological variation were anticipated *a priori*, the data were processed using TIC normalization and the spectra peak picked according to local maxima. The peaks were aligned across spectra and filtered to remove the peaks that were present in fewer than 1% of locations. This resulted in 160 peaks. Finally, peaks with observed intensities below the boundary of detection were set to the minimum detected intensity for that peak. All the peak intensities were $\log_2$ transformed to match the assumptions of the downstream statistical analysis.

**Statistical analysis.** We evaluated the sensitivity of various statistical analysis approaches with respect to the number of detected differentially abundant analytes. The importance of matching the experimental design was evaluated by comparing the statistical models that either account for, or ignore, the paired nature of the design. The importance of characterizing the existing sources of variation was evaluated by comparing models that either account

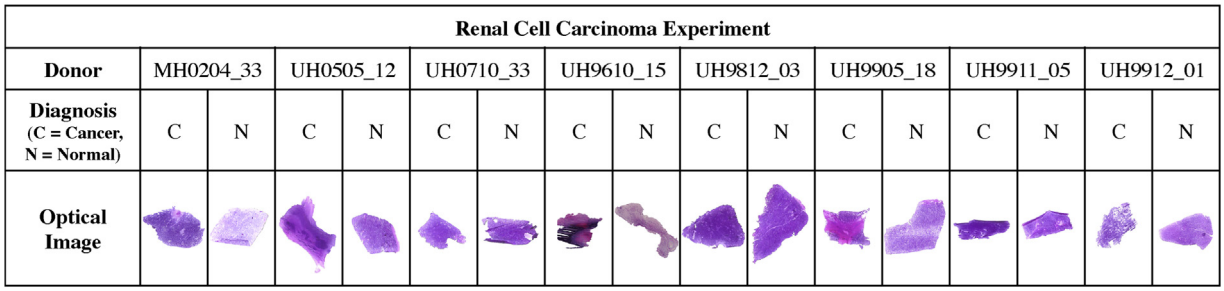


**Fig. 1.** The human renal cell carcinoma (RCC) experiment with paired design. Both "cancer" and "normal" tissues were collected from each of the 8 donors. The images are H & E stained serial sections used for pathological diagnosis.

for within-tissue spatial autocorrelation or summarize the locations in a tissue by averaging $\log_2$ intensities of each analyte in each tissue. Approaches with more differentially abundant analytes were more sensitive.

In addition to the sensitivity, the specificity of the approaches was evaluated by "same-same" comparisons. Locations in each tissue were separated into a top and a bottom half, and compared as if they were different conditions. Approaches with fewer differentially abundant analytes were more specific.

Finally, the importance of matching the research goal was evaluated by comparing the results of hypothesis testing to that of PCA and PLS-DA.

### 3.2. CpG: Nano-DESI MSI of ischemic stroke preconditioning in mouse brain

CpG is an unmethylated oligodeoxynucleotide that has been shown to stimulate the toll-like receptor 9 and induce neuroprotection against ischemic damage, for example ischemic stroke, if administered as a preconditioning agent. The goal of this experiment was to elucidate chemical effects on the brain after CpG administration to mice, and to further understand the molecular mechanisms of this neuroprotection.

**Statistical experimental design.** The experiment was designed as a group comparison (i.e., each condition was represented by different mice) with subsampling (i.e. multiple serial sections of a same mouse tissue). Samples of brain tissue were harvested from three mice with the saline treatment (mice A, B, and C) and three mice with the CpG treatment (mice X, Y, and Z). Three serial sections were gathered from each mouse brain. Mouse Y from the CpG group produced only two sections. The dataset consisted of 17 tissues total (Fig. 2).

**Data acquisition.** Serial tissue sections from the same mouse were placed on the same slide for spectral acquisition. The dataset was acquired in imaging mode, with three micromolar of a lipid standard lysophosphatidylcholine (LPC) 19:0 (from Avanti Polar lipids) included in the nano-DESI solvent, which consisted of 9:1 methanol:water [25]. A Thermo LTQ-orbitrap instrument was used to acquire the data in positive mode. The velocity of the stage was 40 $\mu$m per second, with a scan rate of approximately 1 scan per second. Due to the use of automated gain control, the scan rate varied for each scan. The lines were spaced by 200 $\mu$m. Assuming 1 scan per second, the pixel size was approximately $40 \times 200$ micrometers. The MSI data were saved in RAW file format, and converted to the NetCDF format by the Xcalibur software.

**Data processing.** Data in the NetCDF format were imported into *Cardinal* v1.10.0 using an in-house R script based on *Cardinal*'s `readImzML` function.

During import the data were binned with bin half-width of 200 parts-per-million (ppm), to balance the mass accuracy with the size of the resulting file. The fixed ppm resulted in variable $m/z$ step sizes, such that in $m/z$ units the bin width was wider for larger $m/z$ values. Spectra were then peak picked with respect to local maxima, resulting in 434 peaks. Peaks with observed intensities below the boundary of detection were set to the minimum detected intensity for that peak. The peak corresponding to the sodium adduct of the LPC standard was located in $m/z$ 560.4 bin, while the potassium adduct peak was in the $m/z$ 576.4 bin.

Locations without the signal, and locations outside the tissue boundary were defined as having zero intensities of the standards (i.e., $m/z$ 560.4 or $m/z$ 576.4), as zero intensities likely corresponded to locations with unsuccessful sample ionization or injection. For example, over 5% of the locations on the first tissue section of mouse A produced no signal. Due to the large number of missing locations, data from tissue sections indicated with an asterisk in Fig. 2 were judged as lower quality overall. We evaluated the impact of including or excluding these sections during the statistical analysis.

We evaluated the impact of normalization on detection of differentially abundant analytes in this dataset, by comparing the normalizations that equalize (1) the TIC, (2) the log-intensities of the standard peak at $m/z$ 560.4, and (3) the log-intensities of the standard peak at $m/z$ 576.4.

**Statistical analysis.** In this dataset, we focused on the importance of characterizing the biological and technological sources of variation. First, we evaluated the importance of biological replication, by comparing the analyses with all the mice, versus the analysis with one mouse per treatment (mice C and Z). Second, we evaluated the importance of quality control and reduction of technical variation, by comparing the analyses with or without tissues with low quality data. Finally, we evaluated the sensitivity of the statistical approaches that either account for within-tissue spatial autocorrelation or summarize the locations by averaging $\log_2$ intensities of each analyte in each tissue.

### 3.3. Methods of statistical analysis for class comparison

Analyses for class comparison specified several statistical models, which emphasized different aspects of variation. Below we denote $Y_{ijkl}$ the $\log_2$ intensity of the analyte in condition $i$ (i.e., $i = 1$ for the control, and $i = 2$ for the treatment), donor/mouse $j$, tissue $k$ and location $l$. For experiments with only one tissue per donor, we set $k = 1$. The models were specified separately for each analyte.

**Models for summaries of locations in a tissue.** The models took as input the average of $\log_2$-intensities of the analyte across all the locations in the tissue $\bar{Y}_{ijk\cdot}$. In experiments with one tissue per donor in each group (such as the RCC experiment, ignoring the paired nature of the design), the variation was described using a linear model [39]:

$$\bar{Y}_{ij1\cdot} = \mu + \alpha_i + \varepsilon_{ij}, \quad \alpha_1 = 0, \quad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2) \tag{1}$$

In the model $\mu$ is the average abundance of the analyte in the control population (i.e., the baseline), $\alpha_i$ is the average deviation of the

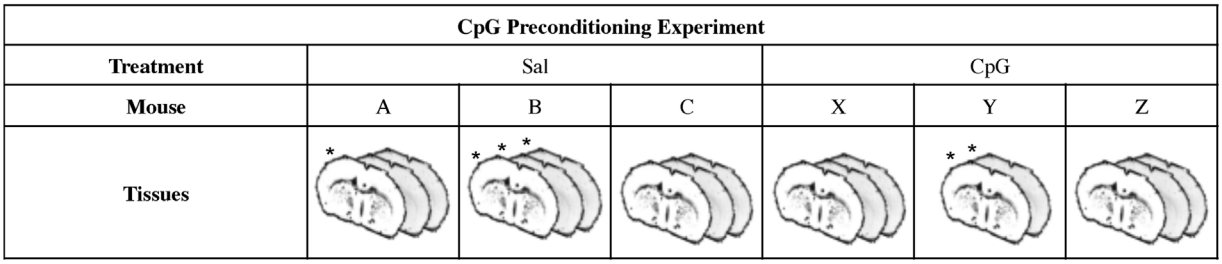


**Fig. 2.** Experimental design of CpG versus Sal MSI dataset. The CpG preconditioning experiment has both biological replicates (multiple mice per treatment) and subsampling (multiple tissues from each mouse). The images schematically illustrate the number of subsamples. The asterisk (*) indicates sections with lower quality data.

analyte abundance between the treatment and the control populations (i.e., the quantity of our main interest), and $\varepsilon_{ij}$ simultaneously represents all the biological and technological sources of variation. Parameters of the model were estimated by Least Squares. The model-based conclusions regarding the differential abundance were identical to those of a two-sample *t*-test with pooled variance.

When the analysis recognized that the experiment included two tissues per donor in a paired design (as in the RCC experiment), it could distinguish the within-subject biological variation and the technological variation. This was accomplished by extending the linear model into a linear mixed effects model with an extra random term $S_j$, which indicates the variation between the individuals $j$ in condition $i$:

$$\begin{aligned} &\bar{Y}_{ij1\cdot} = \mu + \alpha_i + S_j + \varepsilon_{ij1} \\ &\alpha_1 = 0, \quad S_j \overset{iid}{\sim} N(0, \sigma_S^2), \quad \varepsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2) \end{aligned} \tag{2}$$

Parameters of the model were estimated by Maximum Likelihood. The model-based conclusions were also a version of the *t*-test. However, the estimates of biological variation more accurately reflected the experimental design.

In presence of subsamping (as in the CpG experiment), the model similarly extended Eq. (1), but added the notation $j(i)$ to emphasize that tissue donors were *nested* within each condition, and represented multiple tissue sections.

$$\begin{aligned} &\bar{Y}_{ijk\cdot} = \mu + \alpha_i + S_{j(i)} + \varepsilon_{ijk} \\ &\alpha_1 = 0, \quad S_{j(i)} \overset{iid}{\sim} N(0, \sigma_S^2), \quad \varepsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2) \end{aligned} \tag{3}$$

In Eqs. (1)–(3), the models were fit separately for each analyte, and therefore the total number of hypotheses tested was equal to the number of analytes. To adjust for the multiplicity of testing, the False Discovery Rate in the list of differentially abundant analytes was controlled at 5% using the procedure by Benjamini and Hochberg [46].

**Models that account for within-tissue spatial autocorrelation.** Taking as input the log-intensities of analytes at individual locations on the tissue, the models above were extended into Hierarchical Bayesian Spatial models, to distinguish the biological variation within and between the tissues. In experiments with one tissue per donor in each group (such as the RCC experiment, ignoring the paired nature of the design), the Hierarchical Bayesian Spatial Model extended Eq. (1) as follows [43]:

$$Y_{ij1l} = \mu + \alpha_i + S_{j(i)} + \phi_{ij1l} + \epsilon_{ij1l}. \tag{4}$$

The term $\phi_{ij1l}$ is the spatial autocorrelation, which reflects the similarity or heterogeneity of chemical composition in proximal locations, and $\epsilon_{ij1l}$ is the measurement error. The extent of spatial autocorrelation depends on the condition and the subject. It also depends on the ability of the experiment to define homogeneous tissue regions, on the spatial resolution of the MSI, and on whether the individual tissue locations represent one or multiple cell types. The spatial autocorrelation is estimated from the data separately for each analyte.

Similarly, when the analysis recognized that the experiment included two tissues per donor in a paired design (as in the RCC experiment), the Hierarchical Bayesian Spatial Model extended Eq. (2):

$$Y_{ij1l} = \mu + \alpha_i + S_j + \phi_{ij1l} + \epsilon_{ij1l}. \tag{5}$$

In presence of subsamping (as in the CpG experiment), the model in Eq. (3) was similarly extended, while emphasizing that tissue donors were *nested* within each condition, and were represented by multiple tissue sections

$$Y_{ijkl} = \mu + \alpha_i + S_{j(i)} + T_{kj(i)} + \phi_{ijkl} + \epsilon_{ijkl}. \tag{6}$$

Here $T_{kj(i)}$ indicates the variation between tissue sections of a same donor.

Compared to the models for summaries of locations, the introduction of spatial autocorrelation required a fully Bayesian model specification. In particular, for the quantity of our main interest $\alpha_i$,

$$\begin{aligned} &\alpha_1 = 0, \quad \alpha_2 \,|\, \gamma \sim N(0, r(\gamma)\sigma_\alpha^2) \\ &r(1) = 1, \quad r(0) = 0.00001, \quad \gamma \sim Bern(\pi_0) \end{aligned} \tag{7}$$

When population average of the analyte in the treatment group differs from that in the control group, the indicator of differential abundance denoted by $\gamma = 1$, and 0 otherwise. Therefore the hypothesis $H_0$ of no difference in analyte abundance between the conditions, versus the alternative $H_a$ that a difference exists, was tested in terms of posterior probabilities $P(\gamma = 1|\text{data})$. The remaining priors are in A.

Parameters of the model were estimated using Markov Chain Monte Carlo. To adjust the posterior probabilities for the multiplicity of testing, the False Discovery Rate in the list of differentially abundant analytes was controlled at 5% using the procedure by Storey [47]. The datasets and the R code for all the analyses are available upon request.

## 4. Results

We used the two case studies above to evaluate the importance of following the general statistical principles for detection of differentially abundant analytes in MSI experiments. First, we evaluated the importance of accurately reflecting the experimental design.

**Result 1: Recognizing the paired nature of the RCC experiment enhanced the sensitivity of detecting differential abundance.** Table 1 shows that accounting for the paired nature of the experimental design lead to better detection of differentially abundant features between "cancer" and "normal". This is due to the fact that Eqs. (2) and (5) more accurately characterized the biological variation, and viewed each tissue donor as its own control.

The advantage of the paired design is illustrated in Fig. 3 for $m/z$ 821.33. The $\log_2$ intensity of the feature varied substantially

**Table 1**
Number of differentially abundant analytes in the RCC experiment. Ave: Models for summaries of tissue locations. Spat: Models that account for within-tissue spatial autocorrelation. Superscripts in parentheses are equations in Section 3.3 describing the models. "Group comparison": comparison between "cancer" and "normal" tissues. Larger numbers indicate better sensitivity. "Same-same comparison": comparison of two halves of the tissue sections, i.e. sections with a same condition. Smaller numbers indicate better specificity.

| Design | Group comparison | | Same-same comparison | | | |
|---|---|---|---|---|---|---|
| | | | Normal | | Cancer | |
| | Ave | Spat | Ave | Spat | Ave | Spat |
| Unpaired | $13^{(1)}$ | $57^{(4)}$ | $0^{(1)}$ | $0^{(4)}$ | $0^{(1)}$ | $0^{(4)}$ |
| Paired | $36^{(2)}$ | $138^{(5)}$ | $0^{(2)}$ | $9^{(5)}$ | $0^{(2)}$ | $29^{(5)}$ |

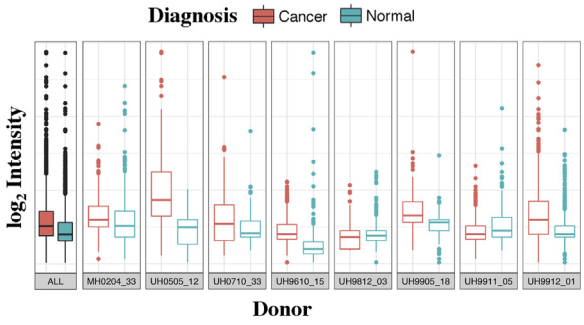


**Fig. 3.** $Log_2$ intensities of $m/z$ 821.33 in the RCC experiment. Filled boxes: all the donors combined. Empty boxes: donor-specific values. Colors indicate the disease. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

**Table 2**
Number of differentially abundant analytes in the CpG experiment. Full: using all tissues in the experiment. Reduced: excluding tissue sections starred in Fig. 2. Ave: Models for summaries of tissue locations. Spat: Models that account for within-tissue spatial autocorrelation. Rows in the table correspond to the normalization strategies. Superscripts in parentheses are equations in Section 3.3 describing the models. Larger numbers indicate higher sensitivity.

| Normalization | Full | | Reduced | |
|---|---|---|---|---|
| | $Ave^{(3)}$ | $Spat^{(6)}$ | $Ave^{(3)}$ | $Spat^{(6)}$ |
| TIC | 0 | 10 | 0 | 23 |
| $m/z$ 560.4 | 0 | 7 | 0 | 49 |
| $m/z$ 576.4 | 0 | 7 | 0 | 52 |

between the donors. Combined across all the donors, the $log_2$ intensity of the feature overlapped substantially between the conditions. On the other hand, the shift between tissues of a same donor pointed more prominently to the same direction (the feature was up-regulated in "cancer" at various extents in most of the donors). Combining the within-donor analyses of the feature facilitated the detection of systematic changes.

Next, we evaluated the importance of accounting for all the existing sources of variation.

**Result 2: Recognizing the spatial autocorrelation in the RCC and the CpG experiments enhanced the sensitivity, and did not undermine the specificity, of detecting differential abundance.** For the RCC experiment, Table 1 shows that models recognizing spatial autocorrelation improved the sensitivity of detecting differential abundance, as compared to the models for summaries of tissue locations. This is particularly true for Eq. (5), which accounted for the paired nature of the experimental design. Similarly, for the CpG experiment, Table 2 shows that recognizing spatial autocorrelation improved the sensitivity of detecting differential abundance across the tissue subsets and normalizations.

The improved performance was due to the fact that the Hierarchical Bayesian Spatial Model could extract richer information

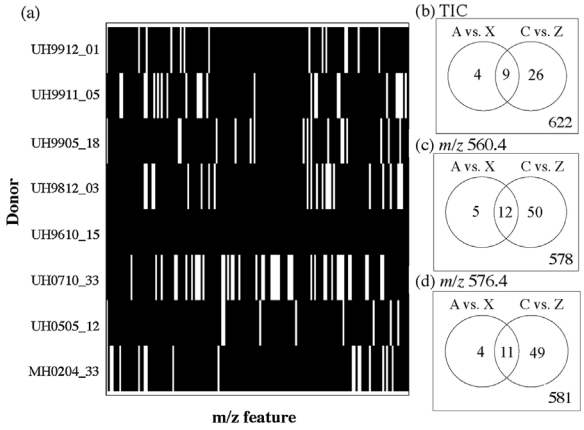


**Fig. 4.** Differentially abundant analytes in unreplicated experiments. (a) Comparisons in single pairs of "cancer" and "normal" tissues in the RCC experiment. *X* axis: 160 $m/z$ features. *Y* axis: tissue donor. Black lines: differentially abundant features for each donor, after fitting Eq. (5). (b) Comparisons in single pairs of mice in the CpG experiment, after the TIC normalization and fitting Eq. (6). Left circle: number of differentially abundant $m/z$ features between mice A (excluding the starred sample in Fig. 2) and mice X. Right circle: number of differentially abundant $m/z$ features between mice C and Z. (c) Same as (b), but with normalization to $m/z$ 560.4. (d) Same as (b), but with normalization to $m/z$ 576.4.

from the tissues in these experiments, which otherwise included a relatively small number of tissue donors (8 in the RCC experiments, and 3 per condition in the CpG experiment). For example, Fig. 3 shows that there was substantial within-tissue variation in $m/z$ 821.33. The models taking as input tissue-wide averages lost this information.

At the same time, the columns "same-same comparison" in Table 1 indicate that the extra model complexity did not substantially undermine the specificity of the results in the RCC dataset. When controlling the FDR at 5% in the set of 160 spectral features, we expected on average 160·0.05=8 false positive discoveries. The results in the "same-same comparisons" among "normal" tissues in Table 1 are comparable with this number. The 'same-same comparisons" among "cancer" tissues exceed the expected number. This may be due to heterogeneity of chemical composition in tumor tissues.

**Result 3: Omitting biological replicates in the RCC and CpG experiments led to overfitting, and to irreproducible results.** In unreplicated experiments, summaries of locations in a tissue contain no information regarding sources of variation. Therefore, models in Eqs. (1)–(3) cannot be applied. While models in Eqs. (4)–(6) are applicable in principle, doing so undermines the reproducibility of the results. For the RCC experiment, Fig. 4(a) shows that detection of differential abundance based on a single tissue donor varied substantially between the donors. More features were found differentially abundant than in Table 1. Similarly, for the CpG experiment, Fig. 4(b) shows little overlap in differentially abundant features in two sets of unreplicated experiments with the TIC normalization, and Fig. 4(c) with the $m/z$ 560.4 normalization. These results are due to the fact that unreplicated experiments do not allow us to assess the extent of between-donor variation of analyte abundance. Therefore, analyses of an unreplicated experiment overfit the patterns of that particular donor, and the results are too sensitive and not necessarily reproducible in another donor.

Next, we evaluated the importance of matching the research goal.

**Result 4: Principal Component Analysis (PCA) and Partial Least Squares Discriminant Analysis (PLS-DA) of the RCC experiment were not successful at determining differentially abundant analytes.** Fig. 5(a) illustrates the result of PCA of tissue locations. The first two principal components only explained 50.2%

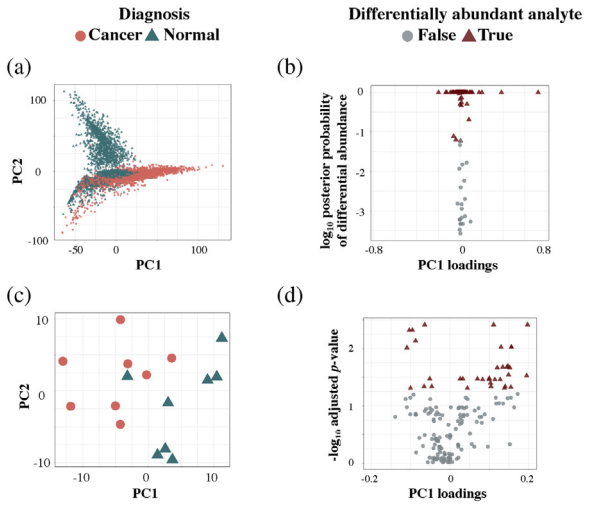


**Fig. 5.** Principal Component Analysis of the RCC experiment. (a) Score plot of tissue locations, in the space of the first two principal components. Each point indicates a location. Locations are colored according to their classification by pathological examination. (b) Loadings of tissue locations in the first principal component in (a), versus estimated posterior probability of differential abundance according to the Hierarchical Bayesian Spatial Model in Eq. (5). Points are spectral features, colored according to differential abundance as determined by the Hierarchical Bayesian Spatial Model, while controlling the FDR at 0.05. (c) Same as (a), but after averaging the $\log_2$ intensities of each analyte over all the locations in a tissue section. (d) Loadings of tissue averages in the first principal component in (c), versus $-\log_{10}$ BH-adjusted $p$-value based on the paired model for averages in Eq. (2). Points are spectral features, colored according to differential abundance determined by the linear model, while controlling the FDR at 0.05. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

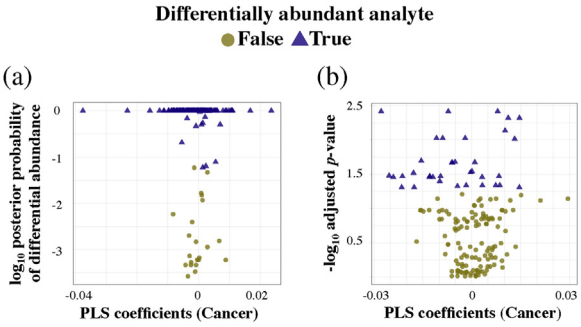


**Fig. 6.** Partial Least Squares Analysis of the RCC experiment. (a) Regression coefficients of PLS-DA, versus estimated posterior probability of differential abundance according to the Hierarchical Bayesian Spatial Model in Eq. (5). Points are spectral features, colored according to their differential abundance, as determined by the Hierarchical Bayesian Spatial Model, while controlling the FDR at 0.05. (b) Regression coefficients of PLS-DA, versus $-\log_{10}$ transformed adjusted $p$-values from Eq. (2). Points are spectral features, colored according to differential abundance as determined by the linear model, while controlling the FDR at 0.05. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

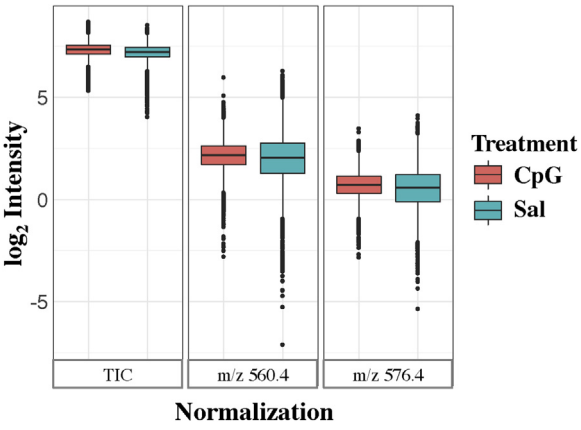


**Fig. 7.** Intensities of $m/z$ 756.513 in all the tissues of the CpG experiment, separated by treatment. Left: after the TIC normalization. Middle: after normalization to the sodium adduct ($m/z$ 560.4). Right: after normalization to the potassium adduct ($m/z$ 576.4).

of the total variation, and did not fully separate the "cancer" and the "normal" locations. The systematic pattern of the points indicated the presence of additional sources of variation, which affected the chemical composition of the locations, but which was not captured by the first two principal components.

Frequently, loadings of the analytes in principal components are interpreted as evidence of their "importance". Fig. 5(b) contrasts the loadings of the first principal component with results of model in Eq. (5). As can be seen, there is little agreement between the two approaches. In particular, class comparison detected differential abundance in many features with small loadings. Moreover, class comparisons controlled the False Discovery Rate in the list of differentially abundant analytes, while the loadings-based approaches of PCA did not.

Fig. 5(c) and (d) repeats the PCA above for averaged $\log_2$ intensities of the analytes in each tissue. The first two principal components only explained 55.8% of the variation of averages. The figures point to the same conclusions as above. Overall, Fig. 5 illustrates that for the RCC experiment, PCA was not successful in detecting differentially abundant analytes.

Fig. 6 shows the results of PLS-DA, which point to the same conclusions as the results of PCA.

Finally, we evaluated the impact of data processing steps.

**Result 5: Excluding tissue sections with poor quality measurements in the CpG experiment improved the sensitivity of detecting differentially abundant analytes.** Table 2 illustrates the benefit of a quality control step preceding the statistical analysis. Discarding tissues with poor quality measurements, based on criteria such as the strength of the signals, increased the sensitivity of detecting differentially abundance with the model that accounts for spatial autocorrelation. The poor quality measurements inflated the estimates of biological and technological variation, and excluding these measurements lead to more sensitive results.

**Result 6: The choices of normalization in the CpG experiment impacted the detection of differentially abundant analytes.** Fig. 7 illustrates that the choice of normalization strongly affected the $\log_2$ intensities of the analytes, and our ability to detect differential abundance. The TIC normalization is frequently used, as it addresses artifacts such as variation in electrospray ionization. However, in the CpG experiment this normalization was possibly affected by high-intensity features with true biological changes in abundance, and equalizing the TIC could over-correct the true variation. Table 2 shows that, as the result, the TIC normalization led to fewer discoveries of differentially abundance in high-quality measurements, and was not effective for this particular experiment. At the same time, normalization with a standard is also fraught with difficulties. Binning can contaminate the standards with signals from other analytes, and different analytes may need to be normalized with different adducts of the standard. For example, in the case of the CpG experiment it may be possible to identify and separately normalize Na and K adducts based on the exact mass difference without identifying the individual molecules.

## 5. Discussion

The case studies confirmed that the general principles of statistical analysis hold for MSI experiments, and should be viewed as guidelines for maximizing the sensitivity and the accuracy of detecting differentially abundant analytes.

First, statistical analysis should match the research question, and use appropriate modeling strategy when the experiment aims at class comparison, as opposed to class discovery or class prediction.

Second, statistical analysis should match the experimental design. Some designs are more effective than others. For example, we have illustrated that for the RCC experiment the paired design is beneficial, as it allows us to view each subject as its own control when comparing the healthy and the cancer tissues. However, the benefit of this design can only be fully exploited if it is followed with the appropriate statistical analysis.

Third, MSI experiments should avoid unreplicated designs, i.e. designs that only focus on tissues from a single donor. Unreplicated experiments limit the scope of conclusions to only that particular donor. As we have seen in the RCC and in the CpG experiments, unreplicated designs lead to overly optimistic and irreproducible results. This is not a surprise because mass spectrometry imaging, as any other measurement technology, does not eliminate the between-subject variation [48].

Forth, in addition to the between-subject variation, statistical analysis should also account for the within tissue spatial autocorrelation. In the case studies in this manuscript, the improvement in sensitivity of the spatial model over averaging was due to appropriately using the individual intensities from all the measured locations. This enhanced the information from an experiments with a relatively small number of donors. However, if the experiment included additional independent tissue donors, the contribution of the within-tissue variation to the sensitivity of the results would likely decrease, and the overall sensitivity of the experiment would likely increase.

Fifth, as we have seen, data processing and quality control strongly affect the results of the statistical analysis. Subjective inclusion and exclusion of measurements, and subjective choices of normalization, can lead to overfitting and to irreproducible results [49]. Similarly, normalization methods and other analysis choices, such as presence or absence of $\log_2$ transformations, or the treatment of zero intensities of the spectral peaks, are also likely to make an impact [50]. To avoid overfitting and maximize the reproducibility of the results, a protocol of quality control and data processing should be specified before conducting the experiment. The protocol can be based on preliminary from a small-scale pilot investigation, and rely on objective criteria of quality (such as the strength of the signal or the amount of missing peaks) and fully automated data analysis steps.

Finally, studies of differential abundance should be followed by additional experimental validation. Two approaches are sometimes considered. First, the same biological samples can be re-analyzed by orthogonal experiments (such as LC–MS) to identify, quantify and interpret the analytes. However, this approach does not constitute validation in a statistical sense, as it confounds the systematic changes induced by the stress or the disease with the biological or sample processing artifacts of the chosen biological replicates. The second approach is to repeat the experiment with new biological specimens, while identifying the analytes. This validation is preferred, and is the gold standard of reproducible research.

Overall, our results demonstrated that appropriate statistical analysis is an important aspect of MSI experiments, and should be carefully considered both before designing the experiment, and during all the data analysis steps.

## Acknowledgments

We thank Allison Dill and R. Graham Cooks of Purdue University for providing access to the Renal Cell Carcinoma experiment. This work was supported by NSF CAREER award DBI-1054826 to O.V, and by funds VR 621-2013-4231 and SSF ICA-6 to I.L. J.L. and M.T. acknowledge support from grant ES024229-01 from the National Institute of Environmental Health Sciences (NIEHS). Part of this research was performed in EMSL, a national scientific user facility sponsored by the DOE's Office of Biological and Environmental Research and located at Pacific Northwest National Laboratory (PNNL) in Richland, WA. PNNL is a multi-program national laboratory operated by Battelle for the DOE under contract DE-AC05-76RLO 1830.

## Appendix A. Priors of Hierarchical Bayesian Spatial Model

The assumptions of the model in Eqs. (4)–(6) are:

$$\alpha_1 = 0, \quad \alpha_2|\gamma \sim N\left(0, r(\gamma)\sigma_\alpha^2\right)$$
$$r(0) = 0.00001, \quad r(1) = 1, \quad \gamma \sim Bern(\pi_0)$$
$$S_{ij}|\sigma_S^2 \sim N\left(0, \sigma_S^2\right), \quad T_{ijk}|\sigma_T^2 \sim N\left(0, \sigma_T^2\right)$$
$$\frac{1}{\sigma_S^2} \sim G(a_S, b_S), \quad \frac{1}{\sigma_T^2} \sim G(a_T, b_T)$$
$$\vec{\phi}_{ijk}|\tau_i^2 \sim ICAR(\tau_i^2, W_{ijk}), \quad \frac{1}{\tau_i^2} \sim G(a_\tau, b_\tau)$$
$$\epsilon_{ijkl}|\sigma_\epsilon^2 \sim N\left(0, \sigma_\epsilon^2\right), \quad \frac{1}{\sigma_\epsilon^2} \sim G(a_\epsilon, b_\epsilon),$$
$$\mu \sim Dirac(c)$$

The intrinsic conditional autoregressive (*ICAR*) model is as described in [51]. It is a conditional model of spatial autocorrelation, with spatial effect $\phi_{ijkl}$ varying around the mean spatial effects at its neighboring locations according to a Normal distribution with variance $\tau_i^2/(\#\text{ofneighborsof}l)$. The neighborhood structures of locations in each tissue are described by the binary matrices $W_{ijk}$.

The hyperparameters of the Gamma distributions were selected as shape $(a_S, a_T, a_\tau, a_\epsilon) = 0.001$ and rate $(b_S, b_T, b_\tau, b_\epsilon) = 0.001$, common choices for a vague prior. The hyperparameter of the condition effect $\alpha_2$ was set to $\sigma_\alpha^2 = 1000$, also forming a vague prior. The prior probability of differential abundance $\pi_0 = 0.1$, reflecting the belief that only a small proportion of features were differentially abundant.

A typical choice of the prior distribution of the baseline condition effect $\mu$ is a non-informative Normal distribution. However, in our experience with MSI experiments, an informative prior of the baseline improved substantially the stability and the convergence of MCMC. Therefore, the implementation in this manuscript specified a Dirac (i.e., point mass) probability distribution, centered on the parameter $c = \sum y_{1jkl}/N$, (i.e., the mean of feature $\log_2$-intensities in the reference condition).

## References

[1] E.R. Amstalden van Hove, D.F. Smith, R.M.A. Heeren, A concise review of mass spectrometry imaging, J. Chromatogr. A 1217 (25) (2010) 3946–3954.
[2] S.R. Ellis, A.L. Bruinen, R.M.A. Heeren, A critical evaluation of the current state-of-the-art in quantitative imaging mass spectrometry, Anal. Bioanal. Chem. 406 (5) (2014) 1275–1289.
[3] A. Römpp, S. Guenther, Y. Schober, O. Schulz, Z. Takats, W. Kummer, B. Spengler, Histology by mass spectrometry: label-free tissue characterization obtained from high-accuracy bioanalytical imaging, Angew. Chem. Int. Ed 49 (22) (2010) 3834–3838.
[4] D. Calligaris, I. Norton, D.R. Feldman, J.L. Ide, I.F. Dunn, L.S. Eberlin, R. Graham Cooks, F.A. Jolesz, A.J. Golby, S. Santagata, N.Y. Agar, Mass spectrometry imaging as a tool for surgical decision-making, J. Mass Spectrom. 48 (11) (2013) 118–1178.
[5] P. Chaurand, D.S. Cornett, P.M. Angel, R.M. Caprioli, From whole-body sections down to cellular level, multiscale imaging of phospholipids by MALDI mass spectrometry, Mol. Cell. Proteomics (2010), mcp-O110.

[6] J.M. Wiseman, D.R. Ifa, Y. Zhu, C.B. Kissinger, N.E. Manicke, P.T. Kissinger, R.G. Cooks, Desorption electrospray ionization mass spectrometry: imaging drugs and metabolites in tissues, Proc. Natl. Acad. Sci. U. S. A. (2008).
[7] B. Prideaux, M. Stoeckli, Mass spectrometry imaging for drug distribution studies, J. Proteomics 75 (16) (2012) 4999–5013.
[8] A. Vegvari, T.E. Fehniger, M. Rezeli, T. Laurell, B. Dome, B. Jansson, C. Welinder, G. Marko-Varga, Experimental models to study drug distributions in tissue using MALDI mass spectrometry imaging, J. Proteome Res. 12 (12) (2013) 5626–5633.
[9] M. Andersson, M.R. Groseclose, A.Y. Deutch, R.M. Caprioli, Imaging mass spectrometry of proteins and peptides: 3D volume reconstruction, Nat. Methods 5 (1) (2008) 101.
[10] T.K. Sinha, S. Khatib-Shahidi, T.E. Yankeelov, K. Mapara, M. Ehtesham, D.S. Cornett, B.M. Dawant, R.M. Caprioli, J.C. Gore, Integrating spatially resolved three-dimensional MALDI IMS with in vivo magnetic resonance imaging, Nat. Methods 5 (1) (2008) 57.
[11] N. Verbeeck, J.M. Spraggins, M.J. Murphy, H. -d. Wang, A.Y. Deutch, R.M. Caprioli, R. Van de Plas, Connecting imaging mass spectrometry and magnetic resonance imaging-based anatomical atlases for automated anatomical interpretation and differential analysis, Biochim. Biophys. Acta (BBA) – Proteins Proteomics 1865 (7) (2017) 967–977.
[12] R.M. Caprioli, T.B. Farmer, J. Gile, Molecular imaging of biological samples: localization of peptides and proteins using MALDI-TOF MS, Anal. Chem. 69 (23) (1997) 4751–4760.
[13] D.R. Ifa, J.M. Wiseman, Q. Song, R.G. Cooks, Development of capabilities for imaging mass spectrometry under ambient conditions with desorption electrospray ionization (DESI), Int. J. Mass Spectrom. 259 (1–3) (2007) 8–15.
[14] J. Laskin, B.S. Heath, P.J. Roach, L. Cazares, O.J. Semmes, Tissue imaging using nanospray desorption electrospray ionization mass spectrometry, Anal. Chem. 84 (1) (2011) 141–148.
[15] E.A. Jones, S.-O. Deininger, P.C.W. Hogendoorn, A.M. Deelder, L.A. McDonnell, Imaging mass spectrometry statistical analysis, J. Proteomics 75 (16) (2012) 4962–4989.
[16] T. Alexandrov, MALDI imaging mass spectrometry: statistical data analysis and current computational challenges, BMC Bioinform. 13 (16) (2012) S11.
[17] M. Morelato, A. Beavis, P. Kirkbride, C. Roux, Forensic applications of desorption electrospray ionisation mass spectrometry (DESI-MS), Forensic Sci. Int. 226 (1–3) (2013) 10–21.
[18] D.N. Correa, J.M. Santos, L.S. Eberlin, M.N. Eberlin, S.F. Teunissen, Forensic chemistry and ambient mass spectrometry: a perfect couple destined for a happy marriage? Anal. Chem. (2016) 2515–2526.
[19] A.L. Oberg, O. Vitek, Statistical design of quantitative mass spectrometry-based proteomic experiments, J. Proteome Res. 8 (5) (2009) 2144–2156.
[20] J.K. Lukowski, E.M. Weaver, A.B. Hummon, Analyzing liposomal drug delivery systems in three-dimensional cell culture models using MALDI imaging mass spectrometry, Anal. Chem. 89 (16) (2017) 8453–8458.
[21] H.E. Hulme, L.M. Meikle, H. Wessel, N. Strittmatter, J. Swales, C. Thomson, A. Nilsson, R.J.B. Nibbs, S. Milling, P.E. Andren, C.L. Mackay, A. Dexter, J. Bunch, R.J.A. Goodwin, R. Burchmore, D.M. Wall, Mass spectrometry imaging identifies palmitoylcarnitine as an immunological mediator during *Salmonella* Typhimurium infection, Sci. Rep. 7 (1) (2017) 2786.
[22] A.L. Dill, L.S. Eberlin, C. Zheng, A.B. Costa, D.R. Ifa, L. Cheng, T.A. Masterson, M.O. Koch, O. Vitek, R.G. Cooks, Multivariate statistical differentiation of renal cell carcinomas based on lipidomic analysis by ambient ionization imaging mass spectrometry, Anal. Bioanal. Chem. 398 (7–8) (2010) 2969–2978.
[23] A.L. Dill, L.S. Eberlin, A.B. Costa, C. Zheng, D.R. Ifa, L. Cheng, T.A. Masterson, M.O. Koch, O. Vitek, R.G. Cooks, Multivariate statistical identification of human bladder carcinomas using ambient ionization imaging mass spectrometry, Chemistry (Weinheim an der Bergstrasse, Germany) 17 (10) (2011) 2897–2902.
[24] J.L. Norris, D.S. Cornett, J.A. Mobley, M. Andersson, E.H. Seeley, P. Chaurand, R.M. Caprioli, Processing MALDI mass spectra to improve mass spectral direct tissue analysis, Int. J. Mass Spectrom. 260 (2–3) (2007) 212–221.
[25] I. Lanekoff, B.S. Heath, A. Liyu, M. Thomas, J.P. Carson, J. Laskin, Automated platform for high-resolution tissue imaging using nanospray desorption electrospray ionization mass spectrometry, Anal. Chem. 84 (19) (2012) 8351–8356.
[26] J.S. Morris, K.R. Coombes, J. Koomen, K.A. Baggerly, R. Kobayashi, Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum, Bioinformatics 21 (9) (2005) 1764–1775.
[27] S.A. Kazmi, S. Ghosh, D.-G. Shin, D.W. Hill, D.F. Grant, Alignment of high resolution mass spectra, development of a heuristic approach for metabolomics Metabolomics 2 (2) (2006) 75–83.
[28] S.O. Deininger, D.S. Cornett, R. Paape, M. Becker, C. Pineau, S. Rauser, A. Walch, E. Wolski, Normalization in MALDI-TOF imaging datasets of proteins: practical considerations, Anal. Bioanal. Chem. 401 (1) (2011) 167–181.
[29] J.M. Fonville, C. Carter, O. Cloarec, J.K. Nicholson, J.C. Lindon, J. Bunch, E. Holmes, Robust data processing and normalization strategy for MALDI mass spectrometric imaging, Anal. Chem. 84 (3) (2012) 1310–1319.
[30] K.A. Veselkov, R. Mirnezami, N. Strittmatter, R.D. Goldin, J. Kinross, A.V.M. Speller, T. Abramov, E.A. Jones, A. Darzi, E. Holmes, J.K. Nicholson, Z. Takats, Chemo-informatic strategy for imaging mass spectrometry-based hyperspectral profiling of lipid signatures in colorectal cancer, Proc. Natl. Acad. Sci. U. S. A. 111 (3) (2014) 1216–1221.
[31] I. Lanekoff, M. Thomas, J.P. Carson, J.N. Smith, C. Timchalk, J. Laskin, Imaging nicotine in rat brain tissue by use of nanospray desorption electrospray ionization mass spectrometry, Anal. Chem. 85 (2) (2013) 882–889.
[32] I. Lanekoff, S.L. Stevens, M.P. Stenzel-Poore, J. Laskin, Matrix effects in biological mass spectrometry imaging: identification and compensation, Analyst 139 (14) (2014) 3528–3532.
[33] C.D. Wijetunge, I. Saeed, B.A. Boughton, J.M. Spraggins, R.M. Caprioli, A. Bacic, U. Roessner, S.K. Halgamuge, EXIMS: an improved data analysis pipeline based on a new peak picking method for EXploring Imaging Mass Spectrometry data, Bioinformatics 31 (19) (2015) 3198–3206.
[34] H.K. Kim, M.L. Reyzer, I.J. Choi, C.G. Kim, H.S. Kim, A. Oshima, O. Chertov, S. Colantonio, R.J. Fisher, J.L. Allen, R.M. Caprioli, J.E. Green, Gastric cancer-specific protein profile identified using endoscopic biopsy samples via MALDI mass spectrometry, J. Proteome Res. 9 (8) (2010) 4123–4130.
[35] S.-O. Deininger, M. Becker, D. Suckau, Tutorial: Multivariate Statistical Treatment of Imaging Data for Clinical Biomarker Discovery, 2010, pp. 385–403.
[36] M. Barker, W. Rayens, Partial least squares for discrimination, J. Chemom. 17 (3) (2003) 166–173.
[37] P. Chaurand, S.A. Schwartz, D. Billheimer, B.J. Xu, A. Crecelius, R.M. Caprioli, Integrating histology and imaging mass spectrometry, Anal. Chem. 76 (4) (2004) 1145–1155.
[38] M. Lipsey, Design Sensitivity: Statistical Power for Experimental Research, 19, Sage, 1990.
[39] M.H. Kutner, C. Nachtsheim, J. Neter, W. Li, Applied Linear Statistical Models, 2005.
[40] A. Cnaan, N. Laird, P. Slasor, Tutorial in biostatistics: using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data, Stat. Med. 16 (20) (1997) 2349–2380.
[41] A. Cassese, S.R. Ellis, N. Ogrinc Potočnik, E. Burgermeister, M. Ebert, A. Walch, A.M.J.M. van den Maagdenberg, L.A. McDonnell, R.M.A. Heeren, B. Balluff, Spatial autocorrelation in mass spectrometry imaging, Anal. Chem. 88 (11) (2016) 5871–5878.
[42] H. Ye, R. Mandal, A. Catherman, P.M. Thomas, N.L. Kelleher, C. Ikonomidou, L. Li, Top-down proteomics with mass spectrometry imaging: a pilot study towards discovery of biomarkers for neurodevelopmental disorders, PLOS ONE 9 (4) (2014) e92831.
[43] A.J. Harry, K.A. Bemis, O. Vitek, Accounting for spatial heterogeneity in design and analysis of mass spectrometry-based imaging experiments, Tech. Rep. (2017).
[44] K.D. Bemis, A. Harry, L.S. Eberlin, C.R. Ferreira, S.M. van de Ven, P. Mallick, M. Stolowitz, O. Vitek, Probabilistic segmentation of mass spectrometry (MS) images helps select important ions and characterize confidence in the resulting segments, Mol. Cell. Proteomics (2016), mcp-O115.
[45] K. Bemis, A. Harry, L.S. Eberlin, C. Ferreira, S.M. van de Ven, P. Mallick, M. Stolowitz, O. Vitek, Cardinal: an R package for statistical analysis of mass spectrometry-based imaging experiments, Bioinformatics 31 (14) (2015) 2418–2420.
[46] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. R. Stat. Soc. Ser. B: Methodol. (1995) 289–300.
[47] J.D. Storey, R. Tibshirani, Statistical significance for genomewide studies, Proc. Natl. Acad. Sci. U. S. A. 100 (16) (2003) 9440–9445.
[48] K.D. Hansen, Z. Wu, R.A. Irizarry, J.T. Leek, Sequencing technology does not eliminate biological variability, Nat. Biotechnol. 29 (7) (2011) 572.
[49] J.P. Simmons, L.D. Nelson, U. Simonsohn, False-positive psychology, Psychol. Sci. 22 (11) (2011) 1359–1366.
[50] S.L. Taylor, G.S. Leiserowitz, K. Kim, Accounting for undetected compounds in statistical analyses of mass spectrometry 'omic studies, Stat. Appl. Genet, Mol. Biol. 12 (6) (2013) 703–722.
[51] S. Banerjee, B.P. Carlin, A.E. Gelfand, Hierarchical Modelling and Analysis for Spatial Data, 2004.