Statistical Principles Define an Open-Source Computational Workflow for Mass Spectrometry Imaging Experiments with Complex Designs: A Case Study of Osteoarthritis

Ethan B Rogers[1], Sai Srikanth Lakkimsetty[1], Kylie Ariel Bemis[1], Charles A Schurman[2], Peggi Angel[3], Birgit Schilling[2], Olga Vitek[1]

[1]Khoury College of Computer Sciences at Northeastern University, Boston, MA; [2]Buck Institute for Research on Aging, Novato, CA; [3]Department of Pharmacology & Immunology at the Medical University of South Carolina, Charleston, SC

Osteoarthritis study: Biological context and research question informed the study design and data structures.

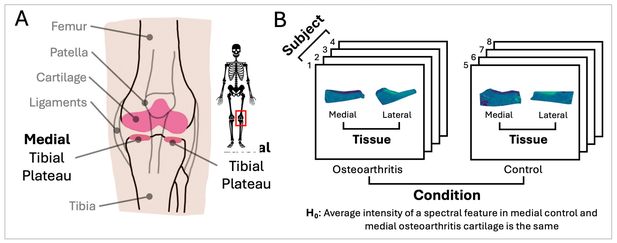


Figure 1: **A:** Anatomy of the knee joint. **B:** Experimental design. Four control subjects and four subjects with osteoarthritis contributed lateral and medial tissues. One null hypothesis of interest is stated below.

Osteoarthritis study: Peak-picking and normalization removed background noise and reduced the overall number of spectral features.

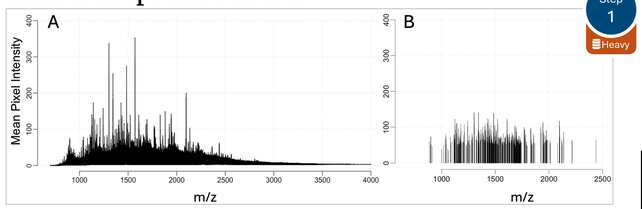
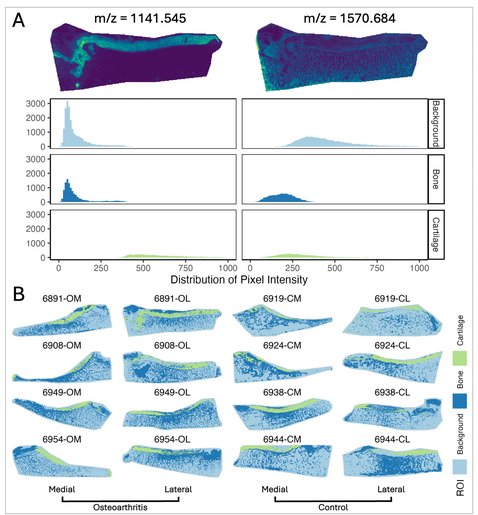


Figure 2.1: **A:** Mean mass spectrum of sample 6891-OL before peak picking and outlier-clipped median TIC normalization. **B:** Mean mass spectrum after peak-picking. Peak-picking reduced the overall number features from over 45,000 to 511. Normalization enabled comparison between tissues. Peak-picking and normalization were done using Cardinal.



Osteoarthritis study: The choice of representative feature used to select ROI greatly influenced the ROI mean feature abundance.

Figure 2.2 (left): **A:** Ions 1141.545 m/z and 1570.684 m/z from sample 6891-OL with unique spatial distributions correlating to cartilage and background structures respectively, and their pixel intensity distributions in each ROI. **B:** The nested segmentation using spatial DGMM and tissue markers highlighted the sponge-like structures of trabecular bone and the cartilage. Segmentation and visualization used Cardinal.

Osteoarthritis study: Clustering and aggregation of spatially similar features reduced the number of features and comparisons without obscuring their relevance.

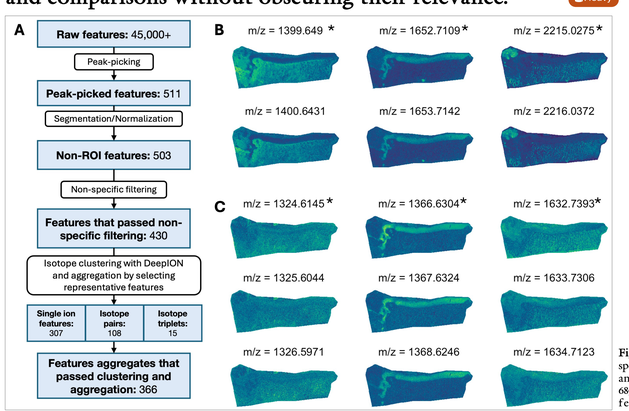


Figure 3 (left): **A:** Workflow of reduction in features from peak picking, segmentation, non-specific filtering, and feature clustering and aggregation. Peak-picking occurred in Cardinal and clustering was done with DeepION. [3] **B:** Isotope groups with two members in tissue 6891-OL had similar spatial distributions. Starred m/z were chosen as the representative feature of the isotope group. **C:** As in **B**, but for isotope groups with three members.

1 Data preprocessing · 2 Filtering and aggregation · 3 Statistical modeling · 4 Statistical inference · 5 Planning future experiments

- **Read data**
  - Formats
  - In/out of memory
- **Peak pick**
  - Alignment, binning & calibration
  - S:N ratio
  - Peak selection
  - Degree of feature reduction
- **Segment ROIs**
  - Based on research question
  - Segmentation method (SSC, sDGMM, HCA, etc.)
  - Choosing segmentation features
  - Minimizing information leakage
- **Normalize**
  - Method (TIC, MIC, RMS)
  - Handling outliers
  - Determining the normalization "constant"
- **Filter, non-specifically**
  - Criteria to isolate noisy features
- **Cluster related features**
  - Clustering method (mass differences, spatial similarity, etc.)
  - Implementation (many choices)
  - Degree of feature reduction
- **Aggregate feature clusters**
  - Mean (large or spatially heterogenous clusters) or representative feature (small clusters with heterogenous expected mean intensity, i.e., isotopes, adducts)
- **Specify statistical model**
  - Choice of model
  - Model features independently
  - Specification of response variable, independent variable(s), fixed and random effects
  - Estimation (SS/ML/REML)
  - Addressing singularity and convergence issues
  - Avoiding common mistakes (subsetting data, pixels as replicates, reporting main effects when interaction is present)
  - Handling missing values, zeros, and left censoring
- **Evaluate model assumptions**
  - Residual plots (quantile-quantile, uniform)
  - Addressing abnormality (transformation)
- **Translate research question to statistical hypothesis**
  - In the terms of the model
- **Tests hypotheses of interest**
  - Calculating standard error
  - Calculating degrees of freedom
  - Modeling choices' impact on variance and degrees of freedom estimates
  - Adjust for multiple testing to control FDR
  - Inspect unadjusted P values for evidence of interdependence
- **Estimate sample size for future work**
  - Get estimates of variance from statistical models
  - Specify anticipated difference, significance threshold, and proportion of differentially abundant features

Step # Resources

- Cardinal or Cardinal data structures
- Standard statistical packages, specific to the experiment
- Other software packages (e.g., DeepION)

Some steps require more memory or computation time – mainly preprocessing and clustering. **Computational intensity is indicated here.**

Introduction

Statistical analysis of multi-tissue mass spectrometry imaging (MSI) experiments with complex designs has the potential to yield meaningful population level conclusions about the relationships between disease states, tissue types, spatial distributions, and ion abundance. However, these data are complex to analyze, with many steps (detailed in the workflow to the *above*) and considerations specific to MSI. Additionally, many tools used to process, visualize, and analyze these data are proprietary – and many open-source alternates are underdeveloped or unready for use by non-computational scientists. Statistical methods for analysis are mature, but complex to understand. Although many are trivial, some steps in these analyses are computationally resource intensive.

We highlight these issues by providing a workflow, replete with recommendations based on experimental and simulated data, and mature, open-source tools and software like Cardinal [1] within which to perform analyses. As a motivating example we demonstrate the steps of this analysis on an MSI dataset of human tibial plateaus from subjects with and without osteoarthritis, illustrating specific decisions using simulations. Details of sample preparation and data acquisition, as well as novel techniques to image bone can be found in [2].

Example R Code.R

```r
library(Cardinal)

# Read Data (Step 1)
m <- readMSIData("./path/to/first/MSI.imzML")
for (i in rest.of.MSI.paths) {
    m <- c(m, readMSIData(i))
}

# Preprocess (Step 1)
m <- m |> normalize() |> smooth() |>
  reduceBaseline() |> peakPick(SNR=7, method="diff") |>
  peakAlign()

# Generate ROIs from known marker (Step 1)
sdgmm <- spatialDGMM(subsetFeatures(m, mz = 1141.55),
                     k = 2)
m$ROI <- factor(sdgmm$class, levels = c(1,2),
                labels = c("cartilage", "other"))

# or generate ROIs manually (Step 1)
cartilage.pixels <- selectROI(m, mz = 1141.55)
m$ROI <- factor(cartilage.pixels, levels = c(T,F),
                labels = c("cartilage", "other"))

# Filter, Non-specifically (Step 2)
m <- summarizeFeatures(m, stat = c("mean", "sd"))
m <- subsetFeatures(m, mean > quantile(mean, .25) |
                       sd > quantile(sd, .25))

# Model and test (Steps 3 & 4)
m$subject <- subject.pixels.factor.vector
m$condition <- condition.pixels.factor.vector
m$tissue <- tissue.pixels.factor.vector
m <- subset(m, ROI == "cartilage")
mt <- meansTest(m, ~ condition * tissue, ~ 1|subject,
                samples = run(all_mse),
                contrast.specs = ~ condition | subtissue,
                contrast.method = "pairwise")
topFeatures(mt)
```

Figure 4: Example code using Cardinal in R for basic untargeted differential analysis. Outputs and code for design, isotope clustering, and sample size calculations are omitted for clarity.

Osteoarthritis study: 1781.857 m/z had the greatest signal to noise ratio of all comparisons in cartilage.

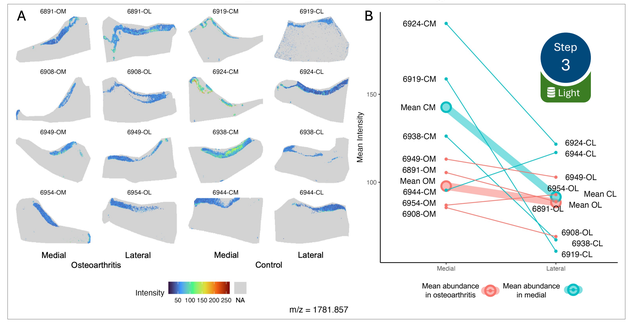


Figure 5: **A:** Single ion images of the bone ROI of each tissue. Missing pixel intensities (*grey*) are dropped during summarization. **B:** Interaction plots of tissue and condition for 1781.857 m/z. Hypothesis A was the difference of medial and lateral osteoarthritis means; points *Mean OM* and *Mean OL*. Hypothesis B was the difference of osteoarthritis and control medial means; points *Mean OM* and *Mean CM*. The difference in means can be thought of as the signal, and the spread of the points as noise; their ratio is the test statistic.

Osteoarthritis study: Within-subjects comparisons had greater signal-to-noise ratios than between-subjects comparisons.

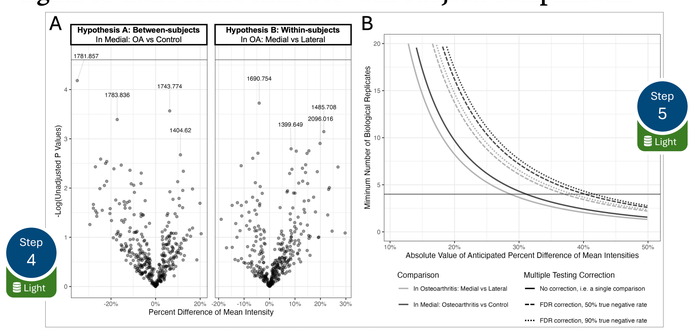


Figure 6: **A:** Volcano plots of comparisons of interest for 366 features. Labeled features have the highest signal-to-noise ratios and lowest p-values. The *grey* line is located at P=0.01. Greater aggregation, or a research question investigating a subset of features would fewer comparisons, increasing power of the experiment. **B:** Minimum number of subjects needed per factor level combination to detect comparison percent difference with based on the osteoarthritis study dataset. Solid horizontal line marks the number of biological replicates in the osteoarthritis study, 4.

Recommendations for designs and analysis of any future MSI experiment with complex designs

Increase signal
- Define specific ROIs using exogenous information
- Use within-subjects comparisons when possible

Increase sample size
- Use previous estimates of variance to determine required sample size

Reduce noise
- Use rigorous QC and standards
- Perform sound normalization and peak-picking
- Cluster and aggregate isotopes and adducts

Reduce number of comparisons
- Reduce tests by clustering and non-specific filtering
- Use targeted interpretation of mass spectra

Use appropriate statistical models
- Specify statistical models that describe all sources of variation
- Avoid double-dipping and using pixels as replicates
- Verify model assumptions
- Correct for multiple testing
- Understand where missing values, outliers, and zeros come from and address them accordingly

[1] Bemis, K. A.; Föll, M. C.; Guo, D.; Lakkimsetty, S. S.; Vitek, O. Cardinal v1.2 A Versatile Open-Source Software for Mass Spectrometry Imaging Analysis. *Nat Methods* **2021**, *20* (12), 1881–1886. https://doi.org/10.1038/s41592-023-02057-7 [2] Schurman, C. A.; Bons, J.; Woo, J. J.; Yee, C.; Tao, N.; Alliston, T.; Angel, P.; Schilling, B. Tissue and Extracellular Matrix Remodeling of the Subchondral Bone during Osteoarthritis of Knee Joints as Revealed by Spatial Mass Spectrometry Imaging. *Preprint* https://doi.org/10.1101/2024.08.05.606482 [3] Guo, L.; Xie, C.; Miao, R.; Xu, J.; Xu, X.; Fang, J.; Wang, X.; Liu, W.; Liao, X.; Wang, J.; Dong, J.; Cai, Z. DeepION: A Deep Learning-Based Low-Dimensional Representation Model of Ion Images for Mass Spectrometry Imaging. *Anal Chem* **2024**, *96* (6), 2529–2536. https://doi.org/10.1021/acs.analchem.3c05076 [4] Guo, D.; Bemis, K.; Rawlins, C.; Agar, J.; Vitek, O. Unsupervised Segmentation of Mass Spectrometric Ion Images Characterizes Morphology of Tissues. *Bioinformatics* **2019**, *35* (14), 2408–2417. https://doi.org/10.1093/bioinformatics/bty1044

The authors declare no competing financial interest.

Simulated study 1: Multivariate segmentation methods that use every feature overfit the data and obscured differential abundance.

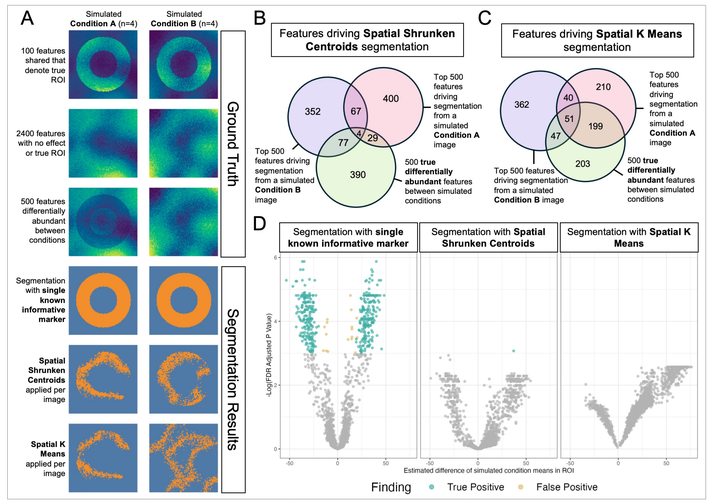


Figure 7.1: **A:** Description of the simulation experimental design and segmentation methods. All simulated images share 100 donut-shaped ROI defining features. **B:** Venn diagrams of top 500 features driving SSC segmentation from a simulated condition A image, a simulated condition B image, and the 500 true differentially abundant features. **C:** The same as B but for SKM. **D:** Volcano plots for t-tests comparing simulated conditions with significance set to FDR-adjusted P<0.05. All segmentation and visualization used Cardinal.

Simulated study 2: Detecting differentially abundant features was more sensitive in within-subject comparisons in than in between-subjects comparisons in the presence of large simulated biological variation.

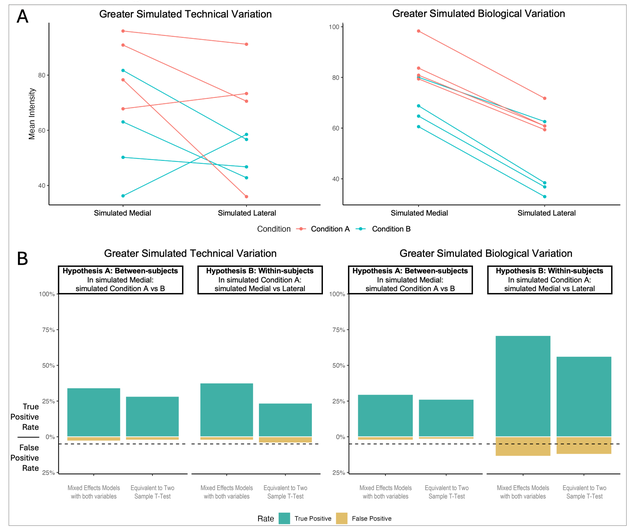


Figure 7.2: **A:** Interaction plots of two representative simulated spectral features. *Left:* simulated within-subject and technical variation larger than the biological between-subject variation. *Right:* simulated within-subject and technical variation smaller than the biological between-subject variation. **B:** True and false positive rates for 300 simulated features analyzed with models in the model table. Within-subject comparisons and mixed effects models were most sensitive. Horizontal dashed line represents a false positive rate of 5%. Comparisons were considered positive when unadjusted P<0.05.

Simulated study 2: Using pixels as biological replicates overfitted the data and produced many false positive differentially abundant features.

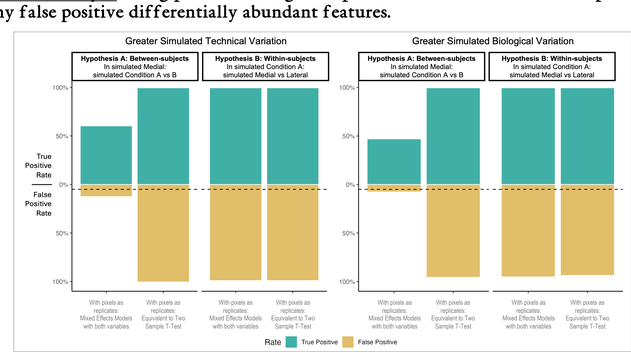


Figure 7.3: Like **7.2B** but using pixels as replicates. *Right:* within-subject and technical variation smaller than the biological between subject variation. When compared to the above, modeling feature intensities in each pixel, and viewing features as replicates, had a much higher rates of false and true positives than models using mean pixel intensity.