Angewandte Chemie
Eine Zeitschrift der Gesellschaft Deutscher Chemiker

Supporting Information

# Morphometric Cell Classification for Single-Cell MALDI-Mass Spectrometry Imaging

*Klára Ščupáková, Frédéric Dewez, Axel K. Walch, Ron M. A. Heeren,* and Benjamin Balluff*

anie_202007315_sm_miscellaneous_information.pdf

## Online methods:
## Samples:

Fresh frozen porcine colon tissue was provided by the MUMC+, Maastricht, The Netherlands, and cryo-sectioned (HM525 Microm, Germany) at 12-µm thickness. The resulting tissue section was placed on an indium tin oxide (ITO)-coated glass slide (Delta Technologies, USA) and stored at -80 °C prior to analysis. Porcine tissue was used in accordance to the Codes of Practice and local regulations in the Central Animal Testing Facilities (CPV) and the University of Maastricht based on the working protocol of Prof. Dr. Nicole D. Bouvy approved by the Animal Welfare Body (AWB) (Instantie voor Dierenwelzijn IvD) under the project license number PV 2017-021 AVD1070020174166.

A 12-µm-thick tissue section of a fresh frozen human diffuse-type gastric carcinoma (DGC) specimen was cut on a cryostat (CM1950, Leica Microsystems, Germany) and placed onto an ITO-coated glass slide (Bruker Daltonik, Germany). The use of the sample was approved by the Institutional Review Board and the Ethics Committee of the Faculty of Medicine of the Technische Universität München with informed patient consent.

**Sample preparation and mass spectrometry imaging experiments:**

2,5-dihydroxybenzoic acid (Sigma Aldrich, USA) and norharmane matrix (Sigma Aldrich) were sublimed at 160 °C for 160 sec using the HTX Sublimator device (HTX Technologies, USA) on the porcine colon and human DGC samples, respectively, for the detection of lipids. Matrix re-crystallization was performed at 50 °C for 1 min using 0.5% ethanol for the porcine colon sample. Water-based Tipp-Ex (BIC S.A., France) fiducial markers were placed on the glass slide, and the slide was scanned at 20× magnification with a desktop slide scanner M8 (PreciPoint, Germany). The resulting optical image is used to define the measurement area. FlexImaging (v5.0, Bruker Daltonik) was used for preliminary co-registration between this optical image and the sample stage coordinate system of the mass spectrometer. Three manually selected control points on distinct fiducial markers, visible both in a live camera image within the mass spectrometer (accessed via the FlexControl software, v5.0, Bruker Daltonik) and the optical image in FlexImaging, were used for this first co-registration.

Matrix-assisted laser desorption/ionization (MALDI) mass spectrometry imaging (MSI) of the porcine colon data was subsequently performed with the timsTOF fleX (Bruker Daltonik) in positive ion mode and in the mass range *m/z* 350–1600. The carcinoma dataset was acquired with the rapifleX MALDI Tissuetyper (Bruker Daltonik) in negative ion reflectron mode, in a mass range between *m/z* 300–1100. For both experiments, the pixel size was set to 10 µm, and 200 spectra were averaged per pixel.

**MALDI-MSI data processing**

Porcine colon dataset: The MALDI-timsTOF fleX data (130,479 pixels, resolving power 40 000 at *m/z* 800) in its SQLite format was loaded into MATLAB R2018b (MathWorks, USA). The spectrum from each pixel was normalized to its root mean square intensity value. The overall mean spectrum from all pixels was calculated and subjected to peak picking using the MATLAB command *mspeaks* (*Bioinformatics* toolbox v2018a) with a minimum absolute peak intensity threshold of 5. This commands returns detected lipid masses, their intensities and peak extensions after default wavelet de-noising. The resulting peak list contained 1252 *m/z* intervals (supplementary figure 5). These intervals were used to reduce the original data by extracting the maximum intensity within the extensions of each interval from every spectrum, resulting in a 130,479 × 1252 data matrix.

Diffuse-type gastric carcinoma dataset: The raw MALDI-TOF rapifleX data (47,355 pixels, resolving power 20 000 at *m/z* 800) was exported as imzML format from FlexImaging (v5.0, Bruker Daltonik) and loaded into MATLAB R2018b (MathWorks). The spectrum from each pixel was normalized to its total ion count. The overall mean spectrum from all pixels was calculated and subjected to peak picking using the MATLAB command *mspeaks* (*Bioinformatics* toolbox v2018a) with a minimum absolute peak intensity threshold of 10E-4. This command returns detected lipid masses, their intensities and peak extensions after wavelet de-noising. The resulting peak list contained 334 *m/z* intervals (supplementary figure 12). These intervals were used to reduce the original data by extracting the maximum intensity within the extensions of each interval from every spectrum, resulting in a 47,355 × 334 data matrix.

**Histological staining & slide scanning**

After MSI experiments, all slides were washed in 100% methanol (Merck KGaA, Germany) to remove the matrix and subjected to conventional hematoxylin & eosin (H&E) histological staining. For this, the slides were submersed for 3 min in each of the series of ethanol (Merck KGaA) solutions: 100%, 100%, 96%, 96% and 70%, 70%. Next, the slides were washed for 3 min in water and submersed for 3 min in hematoxylin (Merck KGaA). Then, the hematoxylin solution was washed off under running water, and the slides were submersed in eosin (Merck KGaA) for 30 sec. The eosin was washed off under running water and the slides were submersed in 100% ethanol for 1 min and 100% xylene (Merck KGaA) for 5 min. Finally, the slides were mounted on glass cover slips using Entellan mounting medium (Merck KGaA) followed by drying for minimally 12 h.

After drying, each slide was scanned by a digital slide scanner (Aperio CS2, Leica, Germany) at 20× magnification. The H&E images are shown in Supplementary Figures 1A and 8A.

**Morphometric image analysis & cell classification**

The scanned high-resolution histological image was loaded into the QuPath software (v0.2.0.m8) where the following steps were carried out:

1. Manual definition of regions of interest (ROIs) for each cell type to be classified:
    a. Porcine colon dataset: 24 ROIs for glandular cells and 17 ROIs for lamina propria cells.
    b. Diffuse-type gastric carcinoma (DGC) dataset: 5 ROIs for tumor cells and 4 ROIs for muscle cells.

2. Automated cell detection in the defined ROIs using the following parameters:

| Parameters | Porcine colon | DGC |
|---|---|---|
| *Detection image* | Hematoxylin OD | Hematoxylin OD |
| *Requested pixel size (µm)* | 0.5 | 0.5 |
| *Background radius (µm)* | 8 | 10 |
| *Median filter radius (µm)* | 0 | 2 |
| *Sigma (µm)* | 1.5 | 1.5 |
| *Minimum area (µm^2)* | 10 | 10 |
| *Maximum area (µm^2)* | 400 | 8000 |
| *Threshold* | 0.1 | 0.051 |
| *Max background intensity* | 2 | 2 |
| *Cell expansion (µm)* | 5 | 5 |
| *Split by shape* | Checked | Unchecked |
| *Include cell nucleus* | Checked | Checked |
| *Smooth boundaries* | Checked | Checked |
| *Make measurements* | Checked | Checked |

3. For both datasets, automatic building of a cell type classifier using Random Forest as classification algorithm with default settings, except:
    a. Use of all extracted morphometric features: Nucleus: Perimeter, Nucleus: Circularity, Nucleus: Max caliper, Nucleus: Min caliper, Nucleus: Eccentricity, Nucleus: Hematoxylin OD mean, Nucleus: Hematoxylin OD sum, Nucleus: Hematoxylin OD std dev, Nucleus: Hematoxylin OD max, Nucleus: Hematoxylin OD min Nucleus: Hematoxylin OD range, Nucleus: Eosin OD mean, Nucleus: Eosin OD sum, Nucleus: Eosin OD std dev, Nucleus: Eosin OD max, Nucleus: Eosin OD min, Nucleus: Eosin OD range, Cell: Area, Cell: Perimeter, Cell: Circularity, Cell: Max caliper, Cell: Min caliper, Cell: Eccentricity, Cell: Hematoxylin OD mean, Cell: Hematoxylin OD std dev, Cell: Hematoxylin OD max, Cell: Hematoxylin OD min, Cell: Eosin OD mean, Cell: Eosin OD std dev, Cell: Eosin OD max, Cell: Eosin OD min, Cytoplasm: Hematoxylin OD mean, Cytoplasm: Hematoxylin OD std dev, Cytoplasm: Hematoxylin OD max, Cytoplasm: Hematoxylin OD min, Cytoplasm: Eosin OD mean, Cytoplasm: Eosin OD std dev, Cytoplasm: Eosin OD max, Cytoplasm: Eosin OD min.
    b. Normalization method: Min-Max
    c. Classes: Balanced
4. Automatic cell detection in the entire specimen
5. Application of the created classifier from 3) to the specimen of interest
6. Export of the classified cell annotations and their coordinates to a text file using QuPath's Groovy-based script editor:

```
import qupath.lib.gui.QuPathGUI

annotations = getDetectionObjects()

def dirResults = QuPathGUI.getSharedDialogHelper().promptForDirectory()
if (dirResults == null)
   return
def fileResults = new File(dirResults, "annotations.txt")

int count = 0
fileResults.withPrintWriter {
               for (pathObject in annotations) {
                              def roi = pathObject.getROI()
                              //print("Annotation: " + pathObject)
                              it.print(pathObject)
                              it.print("\t")
                              pts =  roi.getPolygonPoints()
                              it.print(pts)
                              it.println()
                              count++
               }
}

print 'Done! ' + count + ' result(s) written to ' + fileResults.getAbsolutePath()
```

**Co-registration of MALDI-MSI data to histology image**

The motor positions of all MSI pixels were exported from FlexImaging as a text file and imported into MATLAB R2018b (MathWorks) for every MALDI-MSI dataset. For all subsequent image manipulations, the MATLAB *image* toolbox was required. The control point coordinates of the previous co-registration between sample stage and optical image were extracted from the XML-based FlexImaging file (*.mis) and imported into MATLAB R2018b, resulting in the affine geometric transformation $\Delta_{optical\text{-}MSI}$. Then, the high-resolution, raw H&E image was co-registered to the optical image using at least five manually selected control points based on matchable prominent morphological features of the tissue (MATLAB command *cpselect*), resulting in the affine geometric transformation $\Delta_{HE\text{-}optical}$ (MATLAB command *fitgeotransform)* (Supplementary Figures 2 and 9). Multiplying these two registrations ($\Delta_{HE\text{-}optical}$ * $\Delta_{optical\text{-}MSI}$ = $\Delta_{HE\text{-}MSI)}$ allows directly connecting the MSI data with the H&E image. The H&E image was then cropped to the area of the MALDI-MSI measurement defined by the cropping rectangle [$x_{min}$, $y_{min}$, width, height] (MATLAB command *imcrop)*.

An intensity-based automatic co-registration was used which iteratively transforms and maximizes the similarity between a pair of images (MATLAB command *imregister)* to fine-tune the co-registration between MSI and H&E. The MSI image is a selected *m/z* interval, which can be easily related to a cell type or cellular structures in the H&E image. For the DGC datasets, this was *m/z* 835.55 (PI 34:1), which correlates with the presence of tumor cells, and for the porcine colon dataset, *m/z* 728.53 (Hexosyl Ceramide 36:0 with possible contamination of PS 32:3), which exhibits structures similar to the glandular cells in the mucosa. MSI images underwent hotspot removal by excluding 0.1% of the most intense pixels and thresholding to create binary images to enable an efficient fine co-registration. H&E images were thresholded in their blue channel and image processed including removal of small objects (MATLAB command *bwareaopen*) and smoothing (MATLAB command *imopen*), all to enhance common cell structural details (Supplementary Figures 3–4 and 10–111). The intensity-based, automatic registration between the processed binary images was performed using the multimodal configuration and affine transformation. This registration resulted in a correction of the H&E image, represented as $\Delta_{HE\text{-}HE\_corrected}$.

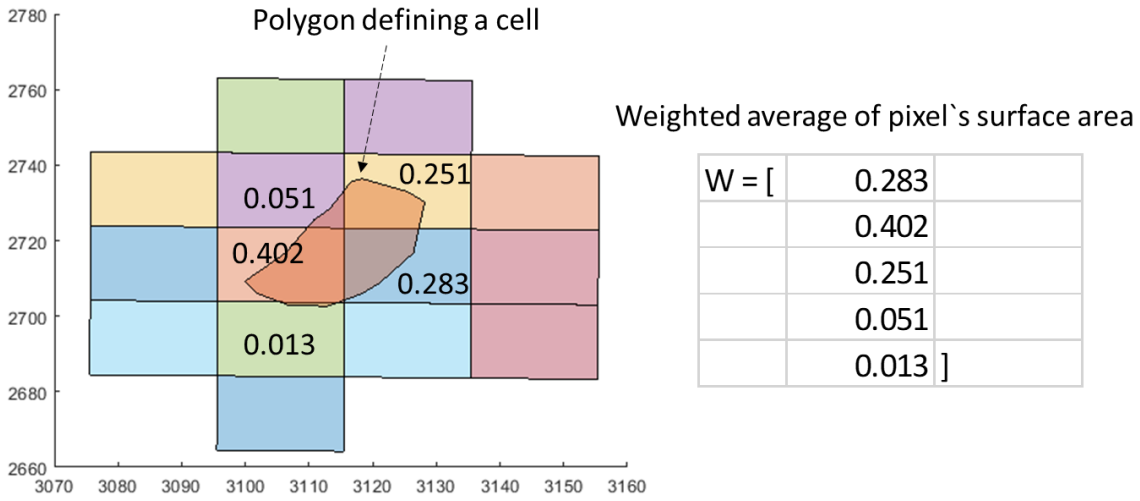
In the porcine colon dataset, the resulting error could be assessed by measuring the Euclidean distance between the H&E pixel positions provided by the geometrical transformation and the actual visible laser burns (Supplementary Figure 6).

**Linking single cells to their molecular profiles**

Once MALDI-MSI and H&E images were aligned, the annotations from QuPath were imported from the previously exported text file. The coordinates of each cell were then transformed by first subtracting the crop points $x_{min}$ and $y_{min}$ and subsequently applying the geometric transformation $\Delta_{HE\text{-}HE\_corrected}$.

To be able to overlay the MSI raster with the cell circumferences, for every MSI pixel $p_i$ with coordinates $p_{i,x}$ and $p_{i,y}$ four point coordinates were created to describe its square area, done by subtracting and/or adding half the MSI pixel size (here 5 µm) to $p_i$: $[(p_{i,x}\text{-}5, p_{i,y}\text{+}5); (p_{i,x}\text{-}5, p_{i,y}\text{-}5); (p_{i,x}\text{+}5, p_{i,y}\text{-}5); (p_{i,x}\text{+}5, p_{i,y}\text{+}5)]$. These MSI square coordinates were transformed to the H&E space by first applying the geometric transformation $\Delta_{HE\text{-}MSI}$ and then by subtracting the crop points $x_{min}$ and $y_{min}$.

Every cell was then checked for its areal overlap with the areas of the 15 closest MSI pixels. The cell's molecular profile was calculated using a weighted-mean across all 15 MSI pixels, where the weights correspond to the degree of overlap between the cell's area and each MSI pixel's area, representing the contribution of each MSI pixel to the cell's molecular profile.




Scheme 1: Visual representation of the weighted average strategy to compute single-cell molecular profiles from several MSI pixels.

**Spatial statistics**

The centroid position of each cell was calculated and used to compute a distance matrix containing the geometric separation between all single cells. This distance matrix was used in the DGC sample to compare the molecular profiles given by MALDI-MSI between muscle cells close to tumor cells and muscle cells far away from tumor cells. We chose to only to take into account and sort distances between muscle cells and tumor cells. An average proximity of a given muscle cell to its *k*-nearest tumor cells was calculated, where *k* was chosen to be 5. These average proximities were thresholded at the first and third quartiles to define muscle cells close to and far away from tumor cells, respectively.

**Correlation analysis of morphometric features with molecular features**

First, a principal component analysis (PCA) was performed in MATLAB R2018b using *statistics* toolbox on the 47,355 pixels × 334 peaks data matrix of the DGC MALDI-MSI dataset. As described above, the principal component scores of the PCA were averaged per cell based on the areal overlap of the cell to the MSI pixels. Next, the morphometric features detected for every cell were exported from QuPath as CSV file and imported to MATLAB. Finally, a Pearson Spearman correlation analysis was performed between the first five principal components of the PCA and the 41 morphometric features.

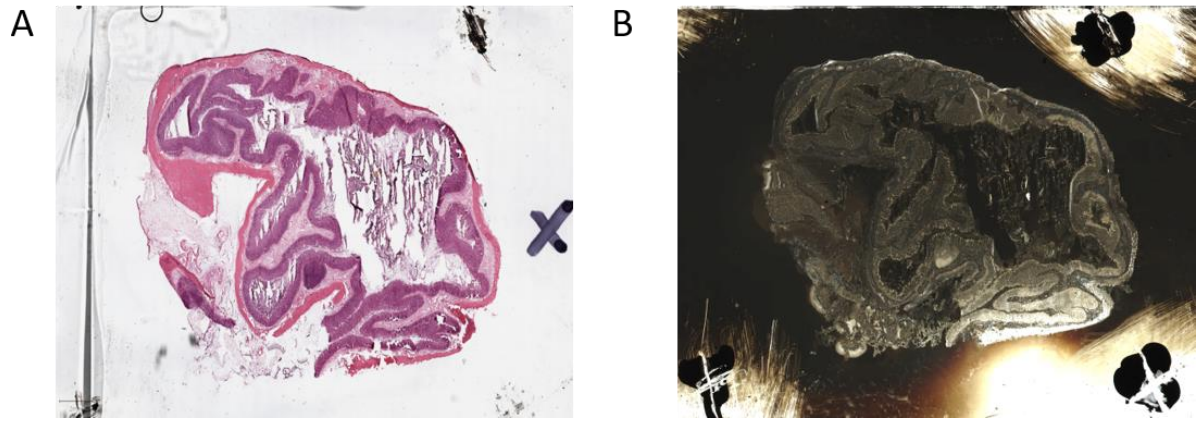
# Morphometric cell classification for single-cell MALDI-MSI

Klára Ščupáková[a], Frédéric Dewez[a, b], Axel K. Walch[c], Ron M.A. Heeren*[a], and Benjamin Balluff[a]

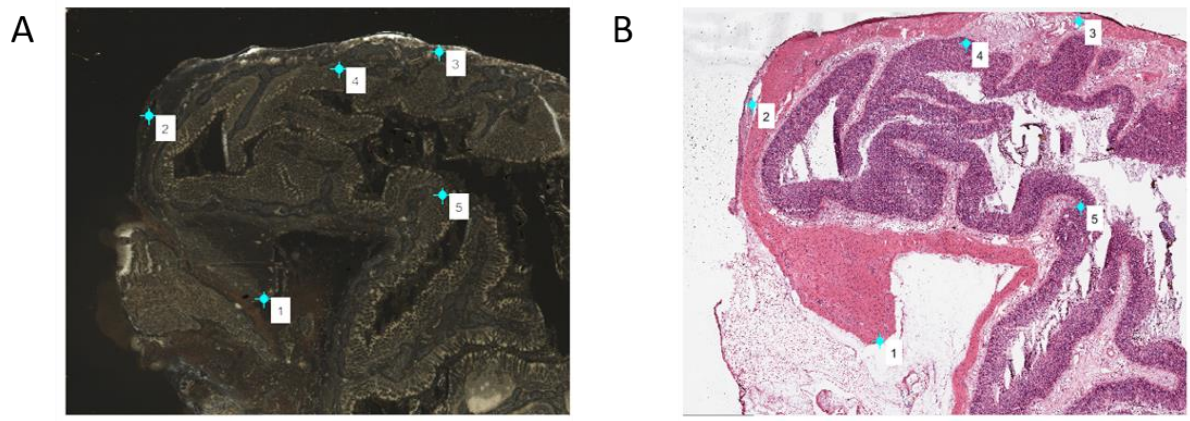
[a] K, Ščupáková; F. Dewez; R.M.A. Heeren; B.Balluff
Maastricht MultiModal Molecular Imaging Institute (M4I)
University of Maastricht
Universiteitssingel 50, 6200 MD Maastricht The Netherlands Department
E-mail: r.heeren@maastrichtuniversity.nl
[b] F. Dewez
Mass Spectrometry Laboratory (MSLab)
University of Liège, Belgium
[c] A.K. Walch
Research Unit Analytical Pathology
Helmholtz Zentrum München,
Oberschleißheim, Germany

## Supplementary figures & tables

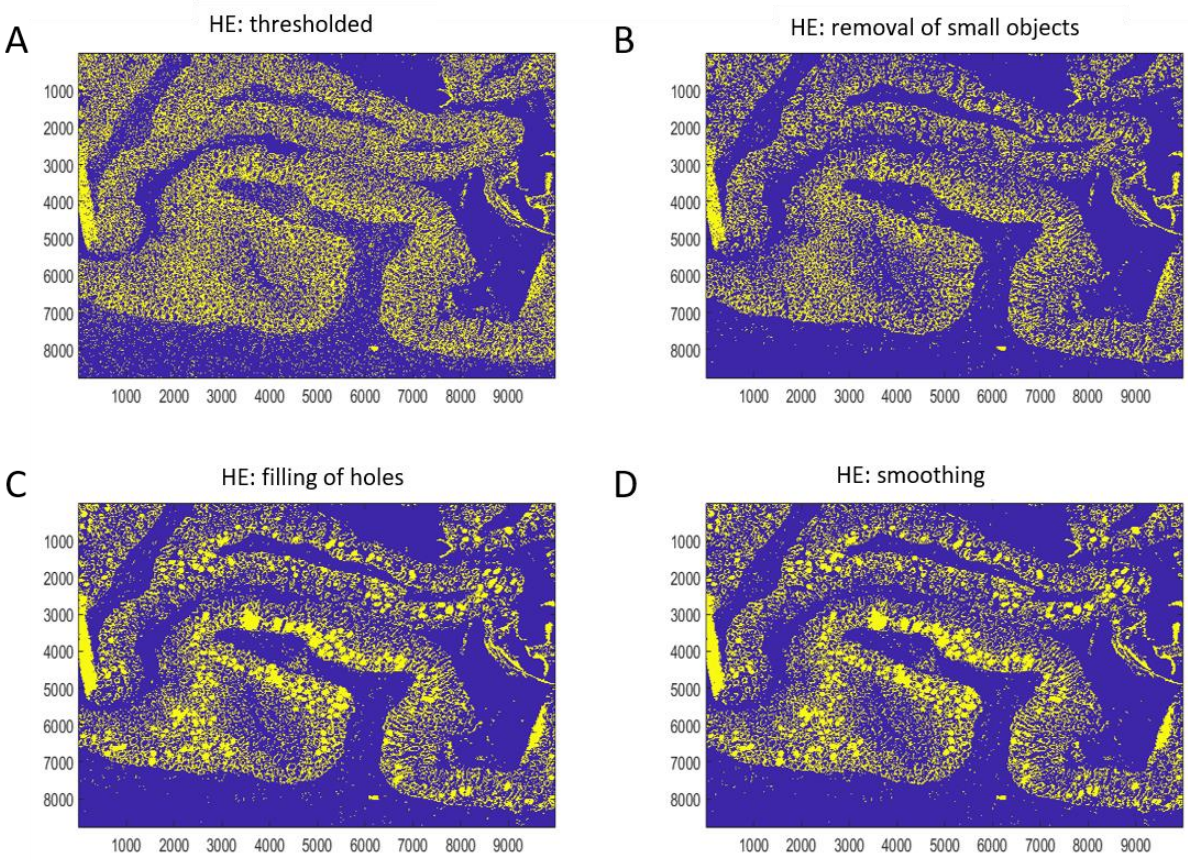
**Dataset #1: Pig colon Tissue**



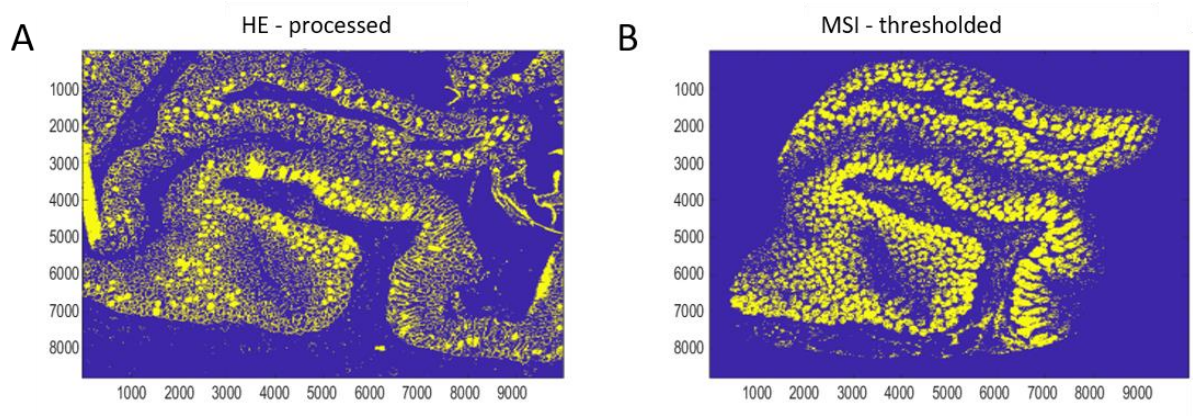

Supplementary figure 1 A: High-resolution hematoxylin and eosin stained image of the entire colon specimen, B: Optical image of the colon tissue acquired prior to MS imaging



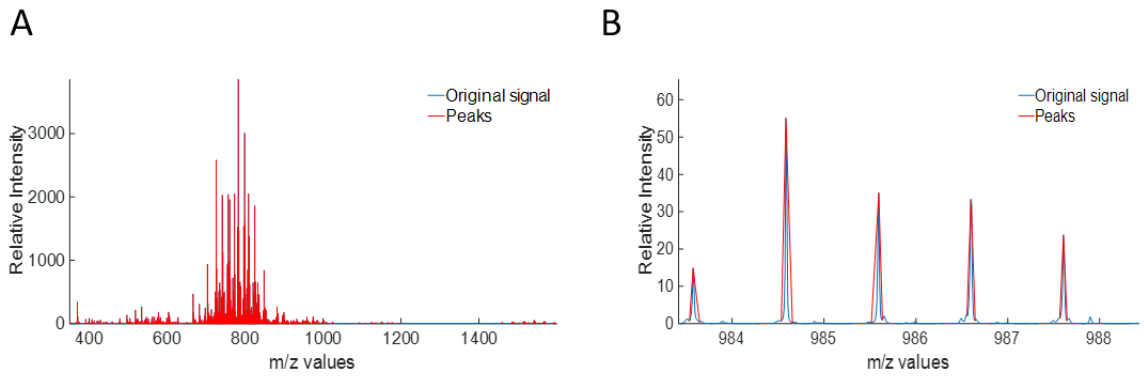

Supplementary figure 2: Initial coarse coregistration based on manual control point selection in both A: optical image and B: hematoxylin and eosin stained image of the same specimen.



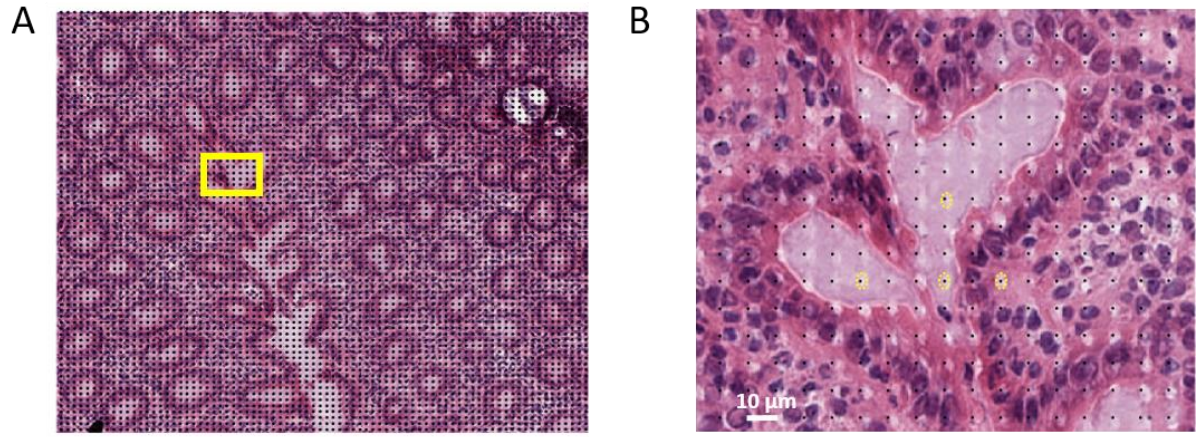

Supplementary figure 3: Preprocessing of the hematoxylin and eosin stained (H&E) image prior to intensity based fine co-registration to increase the comparability between H&E and mass spectrometry imaging (MSI) by matching the resolution and therefore detail-level of MSI. A: Thresholding of the blue channel of the H&E image (hematoxylin stains the nuclei bluish), B: Removal of small objects, C: Filling of empty holes in the H&E image and D: smoothing of the image.



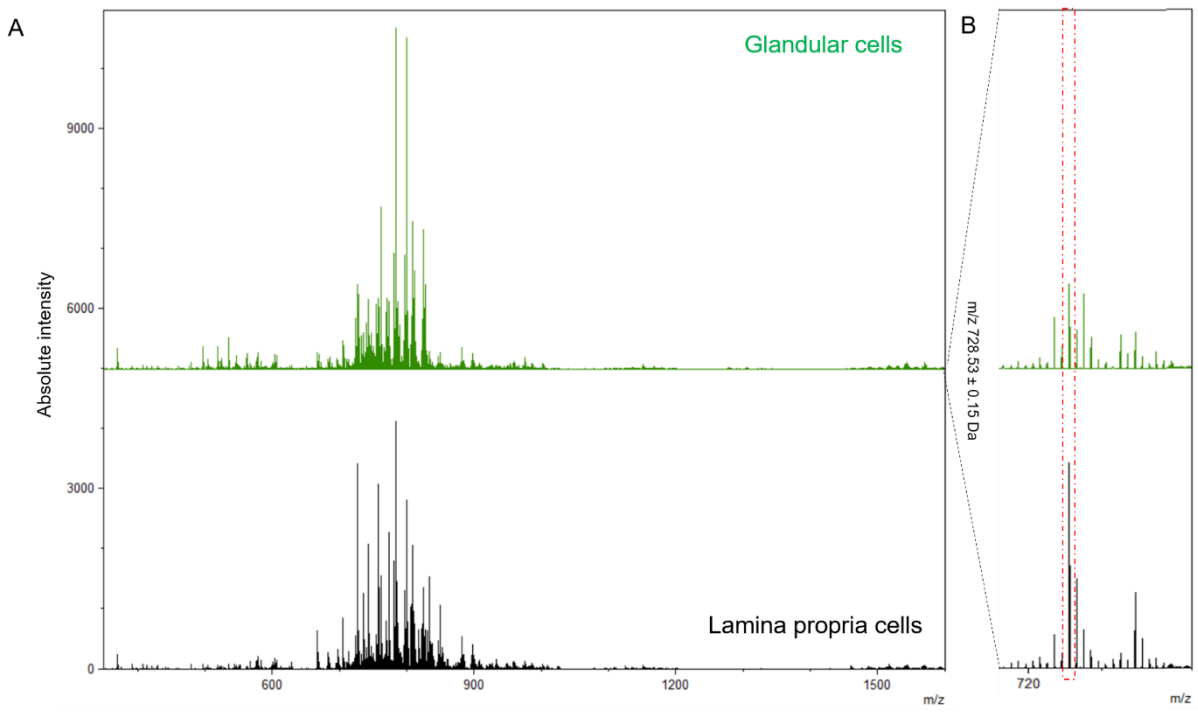

Supplementary figure 4: A: Hematoxylin and eosin stained H&E image and B: Mass spectrometry imaging (MSI) image preprocessed prior to intensity based fine co-registration.




Supplementary figure 5: A: Pig colon tissue mass spectrometry imaging (timsTOF flex) average spectrum. B: Zoom into 984-988 m/z range. Red denotes the peak picked m/z values, blue shows the raw signal.




Supplementary figure 6: Estimation of coregistration error. A: Overlay of H&E image with (denoted in black) centroids of calculated pixels position based on coregistration. B: Zoom of the highlighted area in A, In the H&E image white holes caused by the laser shots are visible. Few of them are highlighted by yellow, dashed circles. In black, the centroids of coregistration-based pixels position are represented. This shows that the coregistration error is well below 1 pixel size.

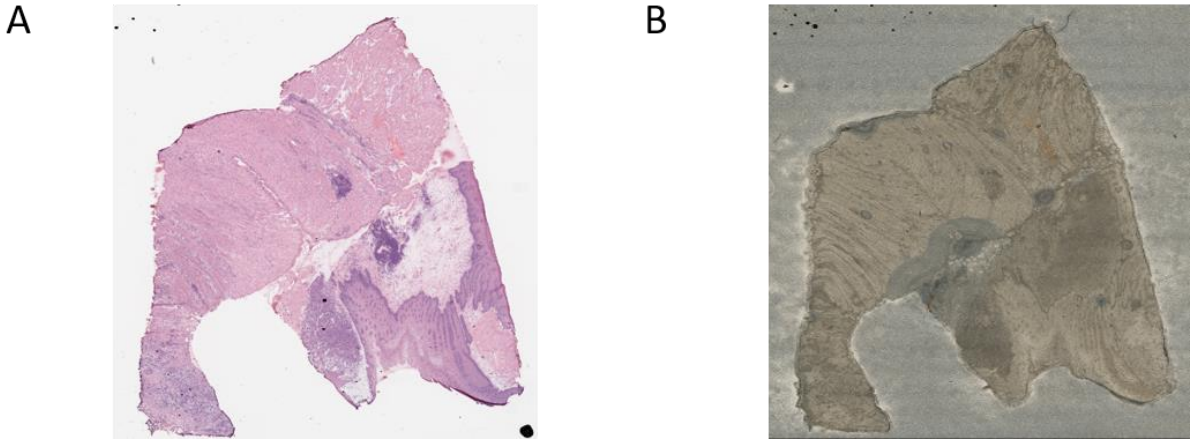



Supplementary figure 7: A: Average spectra from glandular cells (green) and lamina propria cells (black). B: Zoom of A, showing the m/z 728.53 +/- 0.15Da.

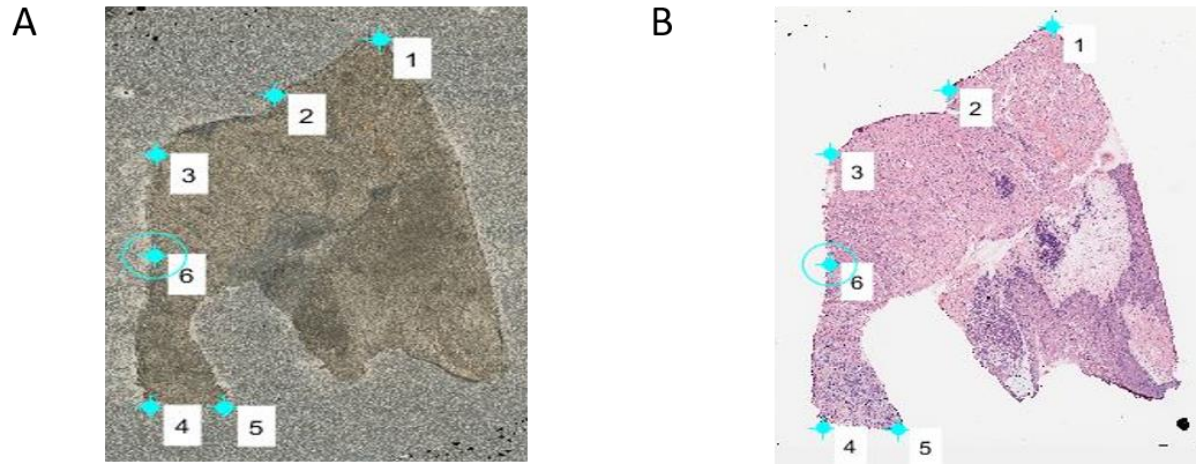
Supplementary table 1: Estimation of coregistration error. MATLAB command imdistline was used to retrieve the Euclidean distance in the hematoxylin and eosin stained (H&E) image between laser marks and calculated pixel position from 27 areas across the measurement region. Next, average distance and standard deviation were calculated. Finally, they were recalculated to micrometers.

| Error [pixels] | Region |
|---|---|
| 2.55 | Northwest |
| 3.77 | Northwest |
| 7.95 | Northwest |
| 7.31 | North |
| 6.73 | North |
| 11.10 | North |
| 6.66 | Northeast |
| 5.95 | Northeast |
| 10.43 | Northeast |
| 9.53 | East |
| 7.59 | East |
| 11.77 | East |
| 10.52 | Southeast |
| 10.15 | Southeast |
| 13.34 | Southeast |
| 6.83 | South |
| 5.69 | South |
| 7.31 | South |
| 11.30 | Southwest |
| 12.95 | Southwest |
| 9.72 | Southwest |
| 8.27 | West |
| 7.16 | West |
| 5.29 | West |
| 8.17 | Center |
| 11.20 | Center |
| 6.88 | Center |
| 8.37 | Average [pixels] |
| 2.69 | Standard deviation [pixels] |
| Size of pixel in um | 0.537 |
| 4.50 | Average [um] |
| 1.45 | Standard deviation [um] |

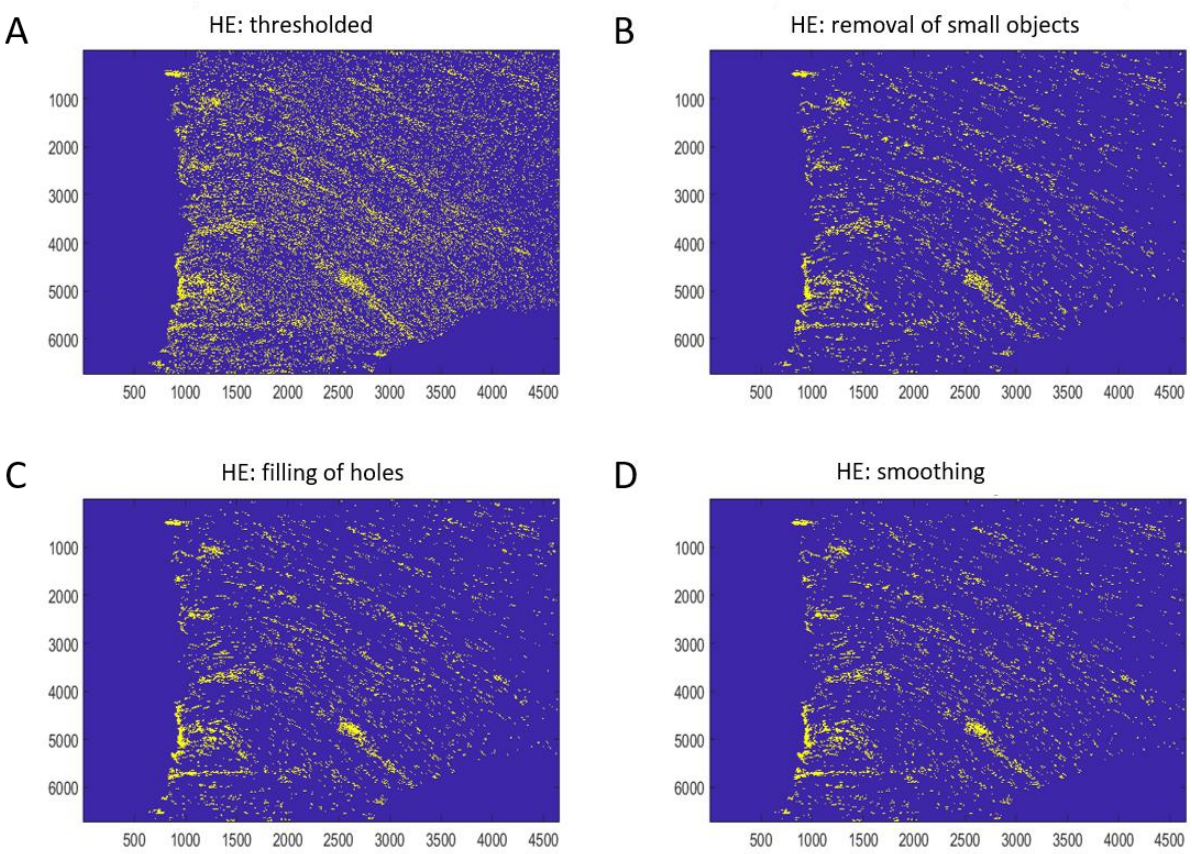
**Dataset #2: Diffuse-type gastric carcinoma**



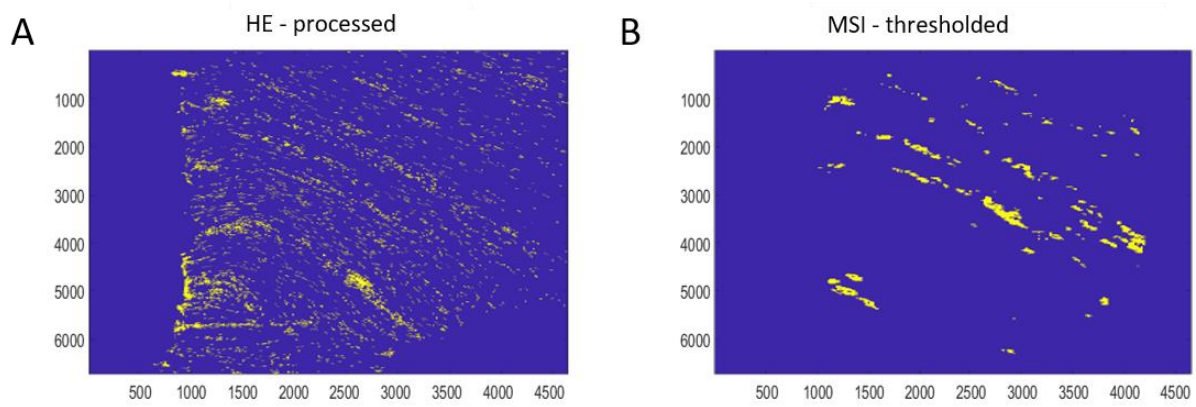

Supplementary figure 8: A: High-resolution hematoxylin and eosin (H&E) image of the diffuse-type gastric carcinoma section, B: optical image of the same specimen taken prior to mass spectrometry analysis.



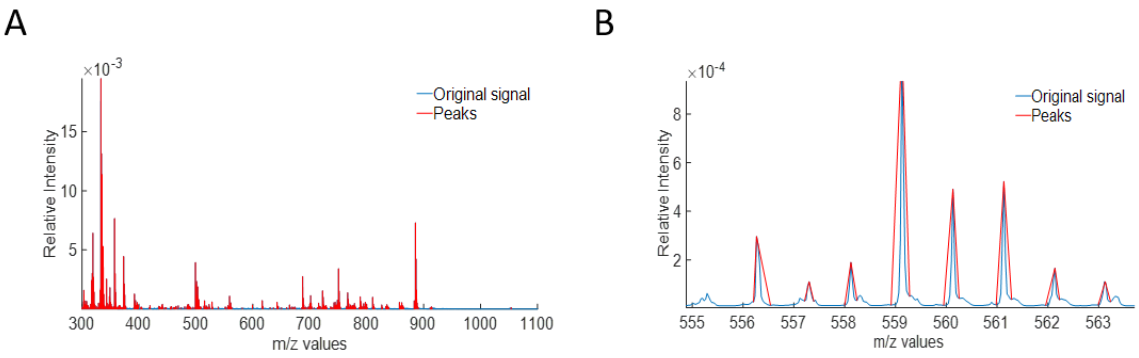

Supplementary figure 9: Initial coarse coregistration based on manual control point selection in both A: optical image and B: hematoxylin and eosin stained image of the same specimen.



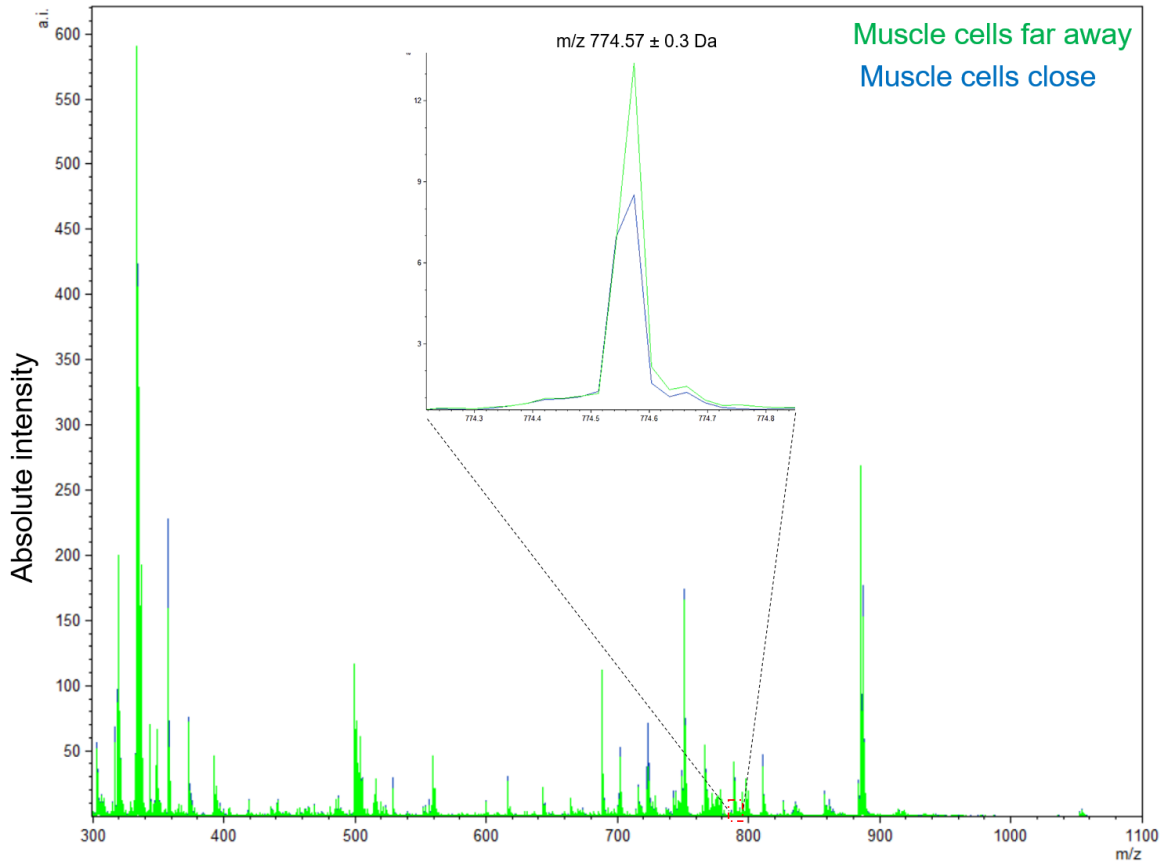

Supplementary figure 10: Preprocessing of the hematoxylin and eosin stained (H&E) image prior to intensity based fine co-registration to increase the comparability between H&E and mass spectrometry imaging (MSI) by matching the resolution and therefore detail-level of MSI. A: Thresholding of the blue channel of the H&E image (hematoxylin stains the nuclei bluish), B: Removal of small objects, C: Filling of empty holes in the H&E image and D: smoothing of the image.




Supplementary figure 11: A: Hematoxylin and eosin stained H&E image and B: Mass spectrometry imaging (MSI) image preprocessed prior to intensity based fine co-registration.




Supplementary figure 12: A: Diffused-type gastric carcinoma tissue mass spectrometry imaging (rapifleX) average spectrum. B: Zoom into 555-563 m/z range. Red denotes the peak picked m/z values, blue shows the raw signal.




Supplementary figure 13: Average spectra from muscle cells far away from tumor cells (green) and muscle cells close to tumor cells (blue). Inset: Zoom of mass range, showing the m/z 774.57 +/- 0.3 Da.