Co-localization features for classification of tumors using mass spectrometry imaging

Paolo Inglese, Gonçalo Correia*, Pamela Pruski*, Robert C Glen, Zoltan Takats

Computational and Systems Medicine, Surgery and Cancer,

Imperial College London, London (UK)

*Contributed equally

Correspondence to: p.inglese14@imperial.ac.uk, z.takats@imperial.ac.uk

Abstract

Statistical modeling of mass spectrometry imaging (MSI) data is a crucial component for the understanding of the molecular characteristics of cancerous tissues. Quantification of the abundances of metabolites or batch effect between multiple spectral acquisitions represents only a few of the challenges associated with this type of data analysis. Here we introduce a method based on ion co-localization features that allows the classification of whole tissue specimens using MSI data, which overcomes the possible batch effect issues and generates data-driven hypotheses on the underlying mechanisms associated with the different classes of analyzed samples.

INTRODUCTION

Mass spectrometry imaging (MSI) has been used to investigate the spatial distributions of molecular species within tissue specimens. In particular, it has shown promising results to serve as a technological platform for the study of chemical heterogeneity associated with the samples of interest. In the particular case of cancer tissue, established protocols exist that are tailored for the imaging analysis of human biopsy specimens (either freshly prepared - frozen tissue - or - formalin fixed paraffin embedded (FFPE), etc. - (Rubakhin *et al.*, 2005; Patel, 2007)) using either desorption electrospray ionization (DESI) (Takáts, Wiseman and Cooks, 2005; Justin M. Wiseman *et al.*,

2008), matrix-assisted laser desorption mass spectrometry (MALDI-MS) (Hillenkamp *et al.*, 1991; Spengler, 1997; Spengler and Hubert, 2002; Cornett *et al.*, 2007; Lemaire *et al.*, 2007; Casadonte and Caprioli, 2011) or secondary ionization mass spectrometry (SIMS) (Colliver *et al.*, 1997; Chandra, 2004; Johansson, 2006; Debois *et al.*, 2009) as the main predominant MSI analytical techniques. MSI spectra obtained in this way have been used to associate ionic species with histological features, to find biochemical differences between healthy or diseased (cancerous, necrotic, etc.) tissue regions or to study the tissue distribution of drugs (J. M. Wiseman *et al.*, 2008; Nilsson *et al.*, 2015; Karlsson and Hanrieder, 2017) and to relate small molecule metabolites and lipids with the expression of protein or other molecular markers (Rawlins *et al.*, 2015). Although some of these methods involve elaborate sample preparation – such as MALDI –, they also suffer from the usual analytical challenges of mass spectrometry-based analysis (ion suppression/matrix effects, intensity drifts, batch effects) (Taylor, Dexter and Bunch, 2018). Furthermore, the inherent complexity of solid tissue analysis brings extra difficulties regarding devising quality control materials and procedures for assessing ionization efficiency and matrix effects, platform stability and intensity drift. For that reason, the capacity to measure and correct the impact of batch and run-order effects is still underdeveloped compared to better established metabolomics and proteomic techniques (Gregori *et al.*, 2012; Fages *et al.*, 2013; Ellis, Bruinen and Heeren, 2014; Goh, Wang and Wong, 2017; Sánchez-Illana *et al.*, 2018).

Although workflows which use peak intensities for classification or regression analyses are widespread in the literature, their effectiveness may be influenced by the magnitude of the batch-effects and other uncontrolled systematic variations. Without any means of the batch-effect correction, the relationship between the true molecular properties of the analyzed samples and the observed and molecular patterns becomes unreliable. Additionally, due to the nature of the MSI

data, combining the information present in a large number of pixel spectra to discriminate entire tissue specimens can be challenging. For instance, studies have shown that MSI data can be used to distinguish different cancer types by selecting a random subset of pixel spectra from a region of interest (Meding *et al.*, 2012), or to determine the risk of metastatic relapse if a certain number of pixel spectra are predicted according to determined labels (Abdelmoula *et al.*, 2016).

A possible alternative approach, employed in the presented work, exploits more extensively the spatial information about the molecular distributions contained in MSI data, representing the molecular information of individual samples by their co-localization patterns. In this way, the whole sample can be expressed in its entirety by a single vector of co-localization features. Additionally, possible issues caused by batch-to-batch variability (Goh, Wang, and Wong, 2017; Bemis *et al.*, 2018) can be mitigated, since co-localization is not affected by systematic variations of ion intensities across multiple samples that can be caused by instrumental or ion source/ion transfer conditions. In general, the biological rationale is that ion co-localization may reflect the underlying biochemical processes that characterize the various classes of samples of interest. A very similar idea has been employed for molecular imaging studies at the tissue and cellular scale, for instance in the case of fluorescence microscopy (Bermudez-Hernandez *et al.*, 2017; Lamberts *et al.*, 2017). Conceptually, the MSI data is converted into a graph, and then network features are used for supervised classification, in an analogous fashion to (Amoroso *et al.*, 2018), where the complex networks framework is employed for classifying patients with neurodegenerative disease. Specifically, Pearson's correlation has been used as a suitable measure for determining the degree of co-localization, in particular in fluorescence microscopy (Manders *et al.*, 1992).

In this paper, we show how the co-localization features can be used to classify three types of cancer specimens, and also provide insight into the possible mechanisms underlying their biochemical differences.

MATERIALS AND METHODS

Sixty-nine specimens from three distinct cohorts of human tumors were cryosectioned and subsequently analyzed by DESI-MSI. The full dataset consisted of twenty-eight breast tumors, part of the dataset published in (Guenther *et al.*, 2015), sixteen colorectal adenocarcinomas, part of the dataset published in (Inglese, Strittmatter, *et al.*, 2018) and twenty-five ovarian cancer samples, part the dataset published in (Dória *et al.*, 2016). All the raw spectra were acquired in the negative ion mode using a custom DESI sprayer ion source interfaced to a high-resolution orbital trapping mass spectrometer [Exactive, Thermo Fisher Scientific] in the range of 150 – 1,000 m/z using the described experimental and instrumental conditions. The mass spectrometry images were therefore split into two random subsets, denoted *cross-validation set* and *test set* consisting of sixteen samples per tumor type, and twelve breast plus nine ovarian cancer samples, respectively.

Pre-processing

All the raw spectral data were first converted into imzML format (centroided mode) (Schramm *et al.*, 2012) and their *m/z* values were re-calibrated using a LOESS model fitted on the distance between the observed and the theoretical *m/z* values of Palmitic acid (255.2330 *m/z*, $[M-H]^-$) (Inglese, Strittmatter, *et al.*, 2018) within a search window of 10 ppm, compatible with the theoretical instrumental error.

The re-calibrated image samples were independently pre-processed through a workflow consisting of a) TIC scaling intensity normalization, b) peak matching, c) log-transformation using MALDIquant for R (Gibb and Strimmer, 2012; Gibb, Franceschi and Gibb, 2017). The choice of

TIC scaling normalization was based on the fact that correlations are insensitive to the domain where the measure is defined. This normalization method allowed us to consider the co-localization of the relative abundances of the ions, reducing the possible effects of systematic variations associated with more quantitative approaches. Peak matching was performed in a way so that no sample could have contributed to the matched *m/z* with more than one peak. No inter-sample normalization was applied to this dataset.

Tissue-unrelated peaks were filtered using the global reference and pixel count filters of the SPUTNIK package for R (Inglese, Correia, *et al.*, 2018). The parameters used are reported in Table 1.

To match the peaks from the different images, peak matching was applied to the average of m/z values corresponding to the non-zero intensity peaks across all the pixels spectra of the *cross-validation set*. Only the peaks detected in all the images were retained for the statistical modeling (Table 2). This strict constraint was necessary to ensure that the different co-localization patterns among the tumor types were involving the molecules observed in all the samples, strengthening their biological significance. The final *common m/z* vector was de-isotoped using the procedure described in (Inglese, Strittmatter, *et al.*, 2018). Manual inspection was performed to identify and remove possible isotopes left after the de-isotoping procedure. The *test set* peaks were matched with the common *m/z* vector of the *cross-validation set* within a search window of ±5 ppm. When multiple peaks matched, the *m/z* value of the most intense peak was selected. If no matched peak was found, the intensity was left equal to 0 in all the pixels. This procedure was intended to evaluate the performances of the classification model with off-sample images.

Co-localization features

A binary ROI representing the area of the image occupied by the tissue was determined by applying k-means clustering to the pre-processed intensity matrix (Inglese, Correia, *et al.*, 2018). After trial-and-error tests, a selection of 4 clusters was used to identify a finer segmentation of the image that could also distinguish those tissue-related areas characterized by low-intensity peaks that would cluster with the off-tissue in case of only 2 clusters. The clusters that were not localized in the corners of the image were merged to define the tissue-related ROI. This rule was determined after the visual inspection of the segmented images (Figure 1).

Spearman's correlations between each pair of matched peaks were calculated within each image sample tissue-related ROI pixels. The asymptotic *t* approximation (Hollander and Wolfe, 1999) was applied to determine the significance of each correlation. The image correlations associated with a Benjamini-Hochberg corrected p-value larger than 0.05 were set equal to 0. Subsequently, the elements of the upper triangular correlation matrix (corresponding to the correlations between different pairs of peaks) were vectorized and used as features for the image sample (Figure 2).

No smoothing or other spatially-related intensity transformations were applied to the ion images before calculating the correlations.

Supervised modeling

The purpose of the supervised model was to predict the tumor type from the co-localization features of an off-sample mass spectrometry image. The procedure consisted of fitting a PLS-DA model on a training set and evaluation of the accuracy of the predicted labels of the validation set. 10-fold cross-validation, repeated 1,000 times, was used to assess the predictive performance of the model, using a varying number of PLS-DA components in a range of 2 to 10.

The accuracies were calculated between the original label vector and the merged predicted labels after a single repetition of the cross-validation procedure. The median accuracy over the 1,000 repetitions was considered as representative of the performances associated with the specific number of components. The maximum average accuracy determined the optimal number of PLS-DA components. A PLS-DA model fitted on the entire *cross-validation set* using the optimal number of components was then used to predict the labels of the *test set*, to assess the robustness of the predictive power of the model.

All the scripts were developed in the R language (R Development Core Team, 2016) and are available at https://github.com/paoloinglese/ion_colocalization.

RESULTS

The pre-processing workflow applied to the *cross-validation set* resulted in 64 common peaks. Among the 21 *test set* samples, one sample had one unmatched peak, two samples had two unmatched peaks, and one sample had three unmatched peaks. The 64 common peaks were annotated by generating the sum formula from the exact mass measurements and by searching the *m/z* values at the online Metlin database (Smith *et al.*, 2005) (Table 3). An *m/z* error threshold 5 ppm was used to accept the putative annotation.

The resulting 2,016 Spearman's correlations were used as features for the PLS-DA models. A number of 7 PLS-DA components corresponded to the maximum average accuracy across the 1,000 repeated 10-fold cross-validations, equal to 0.9205±0.0233 (mean, SD). A comparison with the accuracies obtained after permuting the classification labels confirmed that the observed performances were not due to random chance (Figure 3-A). The confusion matrix of a randomly selected repetition is shown in Figure 3-B.

The PLS-DA model trained on the whole *cross-validation set* was capable of predicting the external *test set* with an accuracy of 0.8095 (Figure 3-C). Although, as expected, the performance dropped, if compared with those of the cross-validation, this result confirmed that the model was adequately capable of distinguishing between tumor types.

Dependence of the correlation values and the size of the ROIs was evaluated. The number of pixels associated with the ROIs varied in a range of [1196, 14345] pixels, with a median of 3726 pixels. The large number of samples (pixels) used to calculate the Spearman's correlations is compatible with previous studies of correlation stability (Schönbrodt and Perugini, 2013). However, since in this case, the ion intensities represent samples drawn from possibly different populations (one per sub-tissue type), a further test was performed to assess the dependency of the observed correlations from the ROI sizes. Firstly, the ROI sizes (in number of pixels) were compared between the three tumor types (Figure 4-A). A Wilcoxon-Mann-Whitney test followed by Benjamini-Hochberg correction between the three pairs of ROI distributions rejected the null hypothesis of equal distribution with a significance level of 95% for the pairs 'breast-colorectal' and 'colorectal-ovarian', but not for the pair 'breast-ovarian'. In particular, breast and ovarian samples had smaller ROIs (median equal to 3205.5 and 3669.5 pixels, respectively) than colorectal samples (median equal to 9,798.5 pixels). Additionally, LOESS models between the number of ROI pixels and each correlation feature of all the *cross-validation set* samples showed that the average correlation values decreased with larger ROIs after (elbow point around 5,000 pixels) (Figure 4-B).

For these reasons, to test whether the classification performances were biased by the ROI size, the same classification procedure was applied to the only breast and ovarian cancer. The

choice of those two classes was based on the fact that they had similar ROI size distributions with a median value lower than the elbow point of the LOESS trends.

The good classification performance of the 1,000 repeated 10-fold cross-validation (median accuracy equal to 0.8750, corresponding to 2 PLS-DA components) (Figure 4-C, D) suggested that the full model performance was not significantly biased by the different ROI sizes of the three tumor types. Analogous results were observed for the *test set* predictions (Figure 4-D).

A multiple pairwise comparison Dunn's test (Dunn, 1964) was employed to determine the significantly different ion-pair correlations between the tumor types. The ten most significant correlations, selected among those with more than a 95% significance level (after Benjamini-Hochberg correction) were further investigated for identifying possible hypotheses on the different biochemical mechanisms associated with the three tumor classes. Molecule pairs correlation significances were calculated as a function of the corrected p-values, defined as $-10\log_{10} p$.

The boxplots of the ten selected correlations are reported in Figure 5. The Cohen's d (Cohen, 1988) confirmed that the selected correlations had a large effect size in each pair of tumors comparisons (Table 4).

The spatial localization of these ions could be investigated by plotting the relative abundance images of the selected correlations. The pairs of ions associated with the five most significant correlations, shown in Figure 6, confirmed that the selected ions were localized in tissue-related areas and did not involve off-tissue regions of the images.

A graph representing the top 10 ion correlations was determined from each pairwise tumor comparison. The median correlations across the samples of the same tumor class were used as the edges of a weighted graph, whereas the peak *m/z* corresponding to the ion pairs were assigned to the nodes. The community structure, determined via short random walks (Pons and Latapy, 2006)

revealed groups of highly correlated ions among those associated with the selected correlation features. Following the assumption behind the presented workflow, these groups of ions can be used to generate hypotheses on the interactions that distinguish the biochemical characteristics of the different cancer types (Figure 7). For this purpose, the elemental composition corresponding to the m/z differences between the ion pairs associated with the selected correlations can provide information about the possible chemical interactions (Table 4).

DISCUSSION

Mass spectrometry imaging is gaining increased interest as a technology capable of mapping the spatial distribution of the molecules of interest in two or three-dimensional samples. Unfortunately, since the quantitative analysis of imaging data is still challenging, and because of the difficulty of controlling the matrix effects hamper its use it as a quantitative tool for the determination of molecular abundances. Ion co-localization analysis, however, can represent a more robust alternative towards more quantitative analysis of this kind of data. Firstly, we showed by applying only a within-sample normalization, that co-localization patterns are not affected by the batch effect. This result was expected because Spearman's correlation detects the linear relationship between two measures independently of their range.

The workflow presented exploits the assumption that local molecular interactions are expressed through co-localization patterns, and that different classes of samples (different tumors in this case) are characterized by diverse metabolic pathways. This hypothesis was corroborated by the high predictive power of the supervised models. Similar performance was observed in an external test set, confirming the robustness of the models.

Additionally, the co-localization patterns can be used for data-driven hypothesis generation, suggesting possible local molecular mechanisms characterizing the samples of interest. The

visualization of the most significantly correlated ions can help to determine regions of interest for further experiments, through data integration with transcriptomics or simply by inspection of the stained samples for histopathological validation, in the case of tissue sections. When represented as a graph the most significant correlations that reveal patterns and clusters of ions that could be used as a part of the hypothesis generation process.

Both positive and negative correlations were considered as feature values for representing the mass spectrometry images. Although a positive correlation may indicate the presence of one or multiple mechanisms locally involving the detected molecules, negative correlations may be seen as reflecting competitive processes like COX1/2, in which Arachidonic Acid is used as a common precursor whereas prostaglandins PG H2 and PG G2 are produced depending on the COX expression. Unfortunately, to date, there is currently not sufficient knowledge to derive such hypotheses exclusively the analysis of mass spectrometry imaging data.

Another limitation of the method is characterized by the representativeness of the studied samples. A single tissue section may be not representative of the complex heterogeneity of the tumor. For this reason, multiple sections or three-dimensional specimens may give a more robust representation of the molecular composition of the analyzed samples.

In the future, various similarity measures will be tested for the quantification of the degree of ion co-localization in cancer tissue and MSI data from different sources. Additionally, biologically-related hypotheses generated by the ion co-localization patterns will be thoroughly investigated through extensive experimental validation, involving transcriptomics and immunohistochemistry measurements.

ACKNOWLEDGEMENTS

We would like to thank Dr. Luisa Doria, Dr. Sabine Guenther, Ms. Anna Mroz, Dr. Nicole Strittmatter for sharing the datasets used in their publications.

This work was supported by Cancer Research UK Grand Challenge Program (Project title: 'A Complete Cartography Through Multiscale Molecular Imaging'). The Department of Surgery and Cancer is funded by grants and infrastructure support by the National Institute for Health Research (NIHR) Imperial Biomedical Research Centre (BRC). GC is supported by the National Institute for Health Research (NIHR) Imperial Biomedical Research Centre (BRC). Funding for P.P. was provided by a European Research Council Consolidator Grant (617896). The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health.

Conflict of Interest: none declared.

REFERENCES

Abdelmoula, W. M. *et al.* (2016) 'Data-driven identification of prognostic tumor subpopulations using spatially mapped t-SNE of mass spectrometry imaging data', *Proceedings of the National Academy of Sciences*, 113(43), pp. 12244–12249. doi: 10.1073/pnas.1510227113.

Amoroso, N. *et al.* (2018) 'Complex networks reveal early MRI markers of Parkinson's disease', *Medical Image Analysis*, 48, pp. 12–24. doi: 10.1016/j.media.2018.05.004.

Bemis, K. A. *et al.* (2018) 'Statistical detection of differentially abundant ions in mass spectrometry-based imaging experiments with complex designs', *International Journal of Mass Spectrometry*. doi: 10.1016/j.ijms.2018.07.006.

Bermudez-Hernandez, K. *et al.* (2017) 'A Method for Quantifying Molecular Interactions

Using Stochastic Modelling and Super-Resolution Microscopy', *Scientific Reports*, 7(1). doi: 10.1038/s41598-017-14922-8.

Casadonte, R. and Caprioli, R. M. (2011) 'Proteomic analysis of formalin-fixed paraffin-embedded tissue by MALDI imaging mass spectrometry', *Nature Protocols*, 6(11), pp. 1695–1709. doi: 10.1038/nprot.2011.388.

Chandra, S. (2004) '3D subcellular SIMS imaging in cryogenically prepared single cells', in *Applied Surface Science*, pp. 467–469. doi: 10.1016/j.apsusc.2004.03.179.

Cohen, J. (1988) 'Statistical power analysis for the behavioral sciences', *Statistical Power Analysis for the Behavioral Sciences*. doi: 10.1234/12345678.

Colliver, T. L. *et al.* (1997) 'Atomic and Molecular Imaging at the Single-Cell Level with TOF-SIMS', *Analytical Chemistry*, 69(13), pp. 2225–2231. doi: 10.1021/ac9701748.

Cornett, D. S. *et al.* (2007) 'MALDI imaging mass spectrometry: Molecular snapshots of biochemical systems', *Nature Methods*, 4(10), pp. 828–833. doi: 10.1038/nmeth1094.

Debois, D. *et al.* (2009) 'In Situ lipidomic analysis of nonalcoholic fatty liver by cluster TOF-SIMS imaging', *Analytical Chemistry*, 81(8), pp. 2823–2831. doi: 10.1021/ac900045m.

Dória, M. L. *et al.* (2016) 'Epithelial ovarian carcinoma diagnosis by desorption electrospray ionization mass spectrometry imaging', *Scientific Reports*. The Author(s), 6(November), pp. 1–11. doi: 10.1038/srep39219.

Dunn, O. J. (1964) 'Multiple Comparisons Using Rank Sums', *Technometrics*, 6(3), pp. 241–252. doi: 10.1080/00401706.1964.10490181.

Ellis, S. R., Bruinen, A. L. and Heeren, R. M. A. (2014) 'A critical evaluation of the current state-of-the-art in quantitative imaging mass spectrometry', *Analytical and Bioanalytical Chemistry*, 406(5), pp. 1275–1289. doi: 10.1007/s00216-013-7478-9.

Fages, A. *et al.* (2013) 'Batch profiling calibration for robust NMR metabonomic data analysis', *Analytical and Bioanalytical Chemistry*, 405(27), pp. 8819–8827. doi: 10.1007/s00216-013-7296-0.

Gibb, S., Franceschi, P. and Gibb, M. S. (2017) 'Package "MALDIquantForeign"'.

Gibb, S. and Strimmer, K. (2012) 'Maldiquant: A versatile R package for the analysis of mass spectrometry data', *Bioinformatics*. Oxford University Press, 28(17), pp. 2270–2271. doi: 10.1093/bioinformatics/bts447.

Goh, W. W. Bin, Wang, W. and Wong, L. (2017) 'Why Batch Effects Matter in Omics Data, and How to Avoid Them', *Trends in Biotechnology*, pp. 498–507. doi: 10.1016/j.tibtech.2017.02.012.

Gregori, J. *et al.* (2012) 'Batch effects correction improves the sensitivity of significance tests in spectral counting-based comparative discovery proteomics', *Journal of Proteomics*, 75(13), pp. 3938–3951. doi: 10.1016/j.jprot.2012.05.005.

Guenther, S. *et al.* (2015) 'Spatially resolved metabolic phenotyping of breast cancer by desorption electrospray ionization mass spectrometry', *Cancer Research*, 75(9), pp. 1828–1837. doi: 10.1158/0008-5472.CAN-14-2258.

Hillenkamp, F. *et al.* (1991) 'Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry of Biopolymers', *Analytical Chemistry*, 63(24). doi: 10.1021/ac00024a716.

Hollander, M. and Wolfe, D. A. (1999) *Nonparametric statistical methods*, *Methods*. doi: 10.1002/.

Inglese, P., Strittmatter, N., *et al.* (2018) 'Network analysis of mass spectrometry imaging data from colorectal cancer identifies key metabolites common to metastatic development.', *bioRxiv*, p. 230052. doi: 10.1101/230052.

Inglese, P., Correia, G., *et al.* (2018) 'SPUTNIK: an R package for filtering of spatially related peaks in mass spectrometry imaging data', *Bioinformatics*. doi: 10.1093/bioinformatics/bty622.

Johansson, B. (2006) 'ToF-SIMS imaging of lipids in cell membranes', *Surface and Interface Analysis*, pp. 1401–1412. doi: 10.1002/sia.2361.

Karlsson, O. and Hanrieder, J. (2017) 'Imaging mass spectrometry in drug development and toxicology', *Archives of Toxicology*, pp. 2283–2294. doi: 10.1007/s00204-016-1905-6.

Lamberts, L. E. *et al.* (2017) 'Tumor-specific uptake of fluorescent bevacizumab-IRDye800CW microdosing in patients with primary breast cancer: A phase I feasibility study', *Clinical Cancer Research*, 23(11), pp. 2730–2741. doi: 10.1158/1078-0432.CCR-16-0437.

Lemaire, R. *et al.* (2007) 'Specific MALDI imaging and profiling for biomarker hunting and validation: Fragment of the 11S proteasome activator complex, reg alpha fragment, is a new potential ovary cancer biomarker', *Journal of Proteome Research*, 6(11), pp. 4127–4134. doi: 10.1021/pr0702722.

Manders, E. M. M. *et al.* (1992) 'Dynamics of three-dimensional replication patterns during the S-phase, analysed by double labelling of DNA and confocal microscopy', *Journal of cell science*, 103, pp. 857–862.

Meding, S. *et al.* (2012) 'Tumor classification of six common cancer types based on proteomic profiling by MALDI imaging', in *Journal of Proteome Research*, pp. 1996–2003. doi: 10.1021/pr200784p.

Nilsson, A. *et al.* (2015) 'Mass Spectrometry Imaging in Drug Development', *Analytical Chemistry*, 87(3), pp. 1437–1455. doi: 10.1021/ac504734s.

Patel, E. (2007) 'Fresh Frozen versus FFPE for Mass Spectrometry Imaging', *Mass*

Spectrometry Reviews. Humana Press, New York, NY, 47(12), pp. 987–92. doi: 10.1002/mas.

Pons, P. and Latapy, M. (2006) 'Computing Communities in Large Networks Using Random Walks', *Journal of Graph Algorithms and Applications*, 10(2), pp. 191–218. doi: 10.7155/jgaa.00124.

R Development Core Team (2016) 'R: A Language and Environment for Statistical Computing', *R Foundation for Statistical Computing Vienna Austria*. doi: 10.1038/sj.hdy.6800737.

Rawlins, C. M. *et al.* (2015) 'Imaging and mapping of tissue constituents at the single-cell level using maldi msi and quantitative laser scanning cytometry', in *Methods in Molecular Biology*, pp. 133–149. doi: 10.1007/978-1-4939-2987-0_10.

Rubakhin, S. S. *et al.* (2005) 'Imaging mass spectrometry: Fundamentals and applications to drug discovery', *Drug Discovery Today*, pp. 823–837. doi: 10.1016/S1359-6446(05)03458-6.

Sánchez-Illana, Á. *et al.* (2018) 'Evaluation of batch effect elimination using quality control replicates in LC-MS metabolite profiling', *Analytica Chimica Acta*, 1019, pp. 38–48. doi: 10.1016/j.aca.2018.02.053.

Schönbrodt, F. D. and Perugini, M. (2013) 'At what sample size do correlations stabilize?', *Journal of Research in Personality*, 47(5), pp. 609–612. doi: 10.1016/j.jrp.2013.05.009.

Schramm, T. *et al.* (2012) 'ImzML - A common data format for the flexible exchange and processing of mass spectrometry imaging data', *Journal of Proteomics*, 75(16), pp. 5106–5110. doi: 10.1016/j.jprot.2012.07.026.

Smith, C. A. *et al.* (2005) 'METLIN: A metabolite mass spectral database', in *Therapeutic Drug Monitoring*, pp. 747–751. doi: 10.1097/01.ftd.0000179845.53213.39.

Spengler, B. (1997) 'Post-source decay analysis in matrix-assisted laser desorption/ionization mass spectrometry of biomolecules', *Journal of Mass Spectrometry*, pp.

1019–1036. doi: 10.1002/(SICI)1096-9888(199711)32:10<1019::AID-JMS595>3.0.CO;2-G.

Spengler, B. and Hubert, M. (2002) 'Scanning microprobe matrix-assisted laser desorption ionization (SMALDI) mass spectrometry: Instrumentation for sub-micrometer resolved LDI and MALDI surface analysis', *Journal of the American Society for Mass Spectrometry*, 13(6), pp. 735–748. doi: 10.1016/S1044-0305(02)00376-8.

Takáts, Z., Wiseman, J. M. and Cooks, R. G. (2005) 'Ambient mass spectrometry using desorption electrospray ionization (DESI): Instrumentation, mechanisms and applications in forensics, chemistry, and biology', *Journal of Mass Spectrometry*, pp. 1261–1275. doi: 10.1002/jms.922.

Taylor, A. J., Dexter, A. and Bunch, J. (2018) 'Exploring Ion Suppression in Mass Spectrometry Imaging of a Heterogeneous Tissue', *Analytical Chemistry*, 90(9), pp. 5637–5645. doi: 10.1021/acs.analchem.7b05005.

Wiseman, J. M. *et al.* (2008) 'Ambient molecular imaging by desorption electrospray ionization mass spectrometry', *Nature Protocols*, 3(3), pp. 517–524. doi: 10.1038/nprot.2008.11.

Wiseman, J. M. *et al.* (2008) 'Desorption electrospray ionization mass spectrometry: Imaging drugs and metabolites in tissues', *Proceedings of the National Academy of Sciences*, 105(47), pp. 18120–18125. doi: 10.1073/pnas.0801066105.

TABLES

Table 1 - SPUTNIK parameters list.

Global similarity filter	
Reference	Binary ROI (k-means)
Method	Pearson's correlation
Threshold	0
Pixel count filter	
Reference	Binary ROI (k-means)
Min. number of connected pixels	9
Aggressiveness	1

Table 2 – MALDIquant parameters used for the peak matching procedure within-sample and between-samples. Min. Freq. The threshold represents the smallest fraction of pixels containing a non-zero intensity used to retain a matched peak.

	Method	**Tolerance**	**Min. Freq. Threshold**
Within sample	strict	10 ppm	0.5%
Between samples	strict	2000 ppm	100%

Table 3 - Annotations of the common m/z values detected after peak matching.

Obs. *m/z*	**Theor. *m/z***	**Error (ppm)**	**Sum formula**	**Adduct**	**Compound**
259.2422	259.2431	3.4	C19H32	M-H-	Nonadecatetraene
279.2321	279.233	3.3	C18H32O2	M-H-	Linoleic acid (C18:2)
303.2321	303.233	2.9	C20H32O2	M-H-	Eicosatetraenoic acid (C20:4)
305.2478	305.2486	2.6	C20H34O2	M-H-	Eicosatrienoic acid (C20:3)
307.2635	307.2643	2.7	C20H36O2	M-H-	Eicosadienoic acid (C20:2)
309.2791	309.2799	2.5	C20H38O2	M-H-	Eicosenoic acid (C20:1)
327.2322	327.233	2.5	C22H32O2	M-H-	Docosatriynoic acid (C22:3)

331.2635	331.2643	2.4	C22H36O2	M-H-	Docosatetraenoic acid (C22:4)
365.3417	365.3425	2.3	C24H46O2	M-H-	Nervonic acid (C24:1)
419.2562	419.2568	1.4	C21H41O6P	M-H-	Cyclic Phosphatidic acid (C18:0)
437.2666	437.2674	1.7	C21H43O7P	M-H-	LysoPA(18:0)
464.3139	464.3146	1.6	C23H48NO6P	M-H-	PE(P-18:0)
480.3089	480.3096	1.4	C23H48NO7P	M-H-	LysoPE(18:0) / LysoPC(15:0)
536.5043	536.5048	0.9	C34H67NO3	M-H-	N-Palmitoylsphingosine / Cer(d34:1)
646.6142	646.6144	0.3	C42H81NO3	M-H-	Ceramide (d42:2)
673.4812	673.4814	0.4	C37H71O8P	M-H-	PA(42:1)
687.5444	687.5446	0.2	C38H77N2O6P	M-H-	SM(d33:1) / PE-Cer(d36:1)
698.513	698.513	0	C39H74NO7P	M-H-	PE(P-34:2) / PE(O-34:3)
699.497	699.497	0	C39H73O8P	M-H-	PA(36:2)
700.5286	700.5287	0.2	C39H76NO7P	M-H-	PE(P-34:1) / PE(O-34:2)
701.5126	701.5127	0.1	C39H75O8P	M-H-	PA(36:1)
714.5079	714.5079	0	C39H74NO8P	M-H-	PE(34:2)
716.5236	716.5236	0	C39H76NO8P	M-H-	PE(34:1)
718.5392	718.5392	0	C39H78NO8P	M-H-	PE(34:0)
722.5131	722.513	-0.1	C41H74NO7P	M-H-	PE(P-36:4) / PE(O-36:5)
723.4968	723.497	0.3	C41H73O8P	M-H-	PA(38:4)
725.5128	725.5127	-0.1	C41H75O8P	M-H-	PA(38:3)
726.5445	726.5443	-0.3	C41H78NO7P	M-H-	PE(P-36:2)
728.5601	728.56	-0.1	C41H80NO7P	M-H-	PE(P-36:1) / PE(O-36:2)
738.5082	738.5079	-0.4	C41H74NO8P	M-H-	PE(36:4)
740.524	740.5236	-0.5	C41H76NO8P / C41H73O8P	M-H-	PE(36:3) / PA(38:4)
742.5394	742.5392	-0.3	C41H78NO8P	M-H-	PE(36:2)
744.5551	744.5549	-0.3	C41H80NO8P	M-H-	PE(36:1)
746.5139	746.513	-1.2	C43H74NO7P	M-H-	PE(P-38:6)
748.53	748.5287	-1.8	C43H76NO7P	M-H-	PE(P-38:5) / PE(O-38:6)
750.5443	750.5443	0.1	C43H78NO7P	M-H-	PE(O-38:5) / PE(P-38:4)
752.5603	752.56	-0.4	C43H80NO7P	M-H-	PC(O-35:4) / PE(P-38:3)
762.5083	762.5079	-0.5	C43H74NO8P	M-H-	PE(38:6)
764.5237	764.5236	-0.1	C43H76NO8P	M-H-	PE(38:5)
766.5393	766.5392	-0.1	C43H78NO8P	M-H-	PE(38:4)

768.5554	768.5549	-0.7	C43H80NO8P	M-H-	PE(38:3)
769.5026	769.5025	-0.2	C42H75O10P	M-H-	PG(36:4)
770.5708	770.5705	-0.4	C43H82NO8P	M-H-	PE(38:2)
771.5183	771.5182	-0.2	C42H77O10P	M-H-	PG(36:3)
772.5867	772.5862	-0.7	C43H84NO8P	M-H-	PE(38:1)
773.5337	773.5338	0.2	C42H79O10P	M-H-	PG(36:2)
774.5472	774.5443	-3.8	C45H78NO7P	M-H-	PE(P-40:6)
776.5619	776.56	-2.4	C45H80NO7P	M-H-	PE(P-40:5) / PE(O-40:6)
778.5761	778.5756	-0.6	C45H82NO7P	M-H-	PE(P-40:4) / PE(O-40:5)
786.5291	786.5291	0	C42H78NO10P	M-H-	PS(36:2)
788.5448	788.5447	-0.2	C42H80NO10P	M-H-	PS(36:1)
792.5553	792.5549	-0.5	C45H80NO8P	M-H-	PE(40:5)
794.5709	794.5705	-0.6	C45H82NO8P	M-H-	PE(40:4)
810.5291	810.5291	0	C44H78NO10P	M-H-	PS(38:4)
812.5454	812.5447	-0.8	C44H80NO10P	M-H-	PS(38:3)
816.5768	816.576	-1	C44H84NO10P	M-H-	PS(38:1)
820.5626	820.5629	0.4	C44H84NO8P	M+Cl-	PC(36:2)
833.5188	833.5186	-0.2	C43H79O13P	M-H-	PI(34:2)
834.5308	834.5291	-2	C46H78NO10P	M-H-	PS(40:6)
857.5186	857.5186	0	C45H79O13P	M-H-	PI(36:4)
859.5352	859.5342	-1.2	C45H81O13P	M-H-	PI(36:3)
861.5501	861.5499	-0.3	C45H83O13P	M-H-	PI(36:2)
883.5344	883.5342	-0.2	C47H81O13P	M-H-	PI(38:5)
885.5501	885.5499	-0.2	C47H83O13P	M-H-	PI(38:4) / Glc-GP(38:4)

Table 4 – Effect size "d" of the ten most significant ion pair correlations in the three pairwise comparisons is reported with the sum formulas corresponding to the m/z difference between the selected ions.

breast-colorectal				
m/z pairs	Δm/z pairs	d	Sum formula	Δ Sum formula
722.5131, 728.5601	6.047	2.575	C41H74NO7P , C41H80NO7P	(+H6 (3 double bonds))
728.5601, 792.5553	63.9952	2.44	C41H80NO7P, C45H80NO8P	(+C4O)
744.5551, 792.5553	48.0002	2.442	C41H80NO8P, C45H80NO8P	(+C4)
812.5454, 816.5768	4.0314	1.968	C44H80NO10P, C44H84NO10P	(+H4 (2 double bonds))
794.5709, 820.5626	25.9917	2.415	C45H82NO8P, C44H84NO8P	(+CH2)
722.5131, 726.5445	4.0314	2.306	C41H74NO7P, C41H78NO7P	(+H4 (2 double bonds))
772.5867, 794.5709	21.9842	2.518	C43H84NO8P, C45H82NO8P	(+C2 -H2)
728.5601, 812.5454	83.9853	2.03	C41H80NO7P, C44H80NO10P	(+C3O3)

303.2321, 728.5601	425.328	2.307	C20H32O2, C41H80NO7P	(+C21H48NO5P)
723.4968, 728.5601	5.0633	2.253	C41H73O8P , C41H80NO7P	(-H7, -O)
breast-ovarian				
m/z pairs	Δm/z pairs	d	Sum formula	Δ Sum formula
722.5131, 788.5448	66.0317	1.984	C41H74NO7P, C42H80NO10P	(+CH6O3)
701.5126, 722.5131	21.0005	1.974	C39H75O8P, C41H74NO7P	(+C2 -OH)
701.5126, 750.5443	49.0317	1.926	C39H75O8P, C43H78NO7P	(+C4H2 -O)
748.53, 857.5186	108.9886	1.971	C43H76NO7P, C45H79O13P	(+C2H3NO6)
722.5131, 752.5603	30.0472	2.099	C41H74NO7P, C43H80NO7P	(+C2H6)
750.5443, 788.5448	38.0005	1.899	C43H78NO7P, C42H80NO10P	(-C / +H2O3)
752.5603, 778.5761	26.0158	1.59	C43H80NO7P, C45H82NO7P	(+C2H2)
687.5444, 885.5501	198.0057	1.88	C38H77N2O6P, C47H83O13P	(+C9H6O7 / -N2)
722.5131, 748.53	26.0169	1.722	C41H74NO7P, C43H76NO7P	(+C2H2)
722.5131, 883.5344	111.0213	1.711	C41H74NO7P, C47H81O13P	(+C6H7O6)
colorectal - ovarian				
m/z pairs	Δm/z pairs	d	Sum formula	Δ Sum formula
834.5308, 861.5501	27.0193	-2.185	C46H78NO10P, C45H83O13P	(-CN /+H5O3)
331.2635, 464.3139	133.0504	-1.978	C22H36O2, C23H48NO7P	(+CH12O5P)
536.5043, 794.5709	258.0666	-1.908	C34H67NO3, C45H82NO8P	(+C11H15O5P)
464.3139, 794.5709	330.257	-1.918	C23H48NO6P, C45H82NO8P	(+C22H35O2)
536.5043, 792.5553	256.051	-1.761	C34H67NO3, C45H80NO8P	(+C2H3O5P)
812.5454, 861.5501	49.0047	-1.908	C44H80NO10P, C45H83O13P	(+CH3O3 /-N)
331.2635, 861.5501	520.2866	-1.779	C22H36O2, C45H83O13P	(+C23H47O11P)
464.3139, 723.4968	259.1829	-1.779	C23H48NO6P, C41H73O8P	(+C18H25O2/ -N)
833.5188, 834.5308	1.012	-1.749	C43H79O13P, C46H78NO10P	(+C3 /-HO3N)
718.5392, 812.5454	94.0062	1.694	C39H78NO8P, C44H80NO10P	(+C5H2O2)

FIGURES

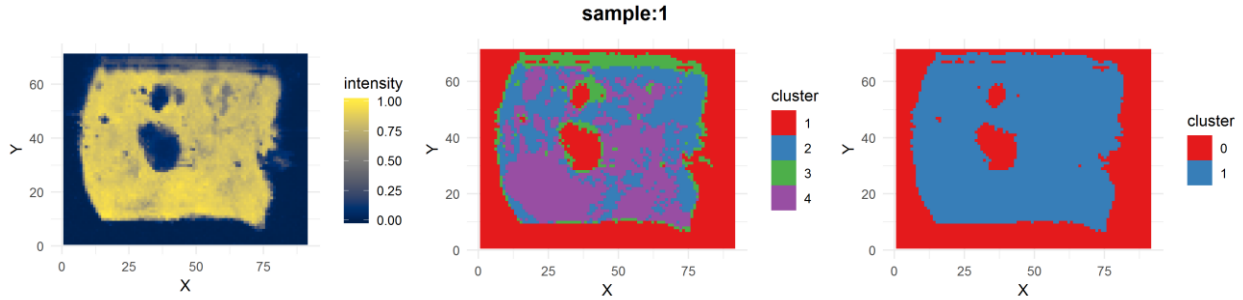

Figure 1 - Tissue-related ROI. K-Means clustering with a number of clusters equal to 4 is applied to the MSI intensity matrix (B). The clusters that are not localized in the off-tissue area (corners used as

reference) are merged into the final tissue-related ROI (C). A comparison with the first principal component scores image (A) shows the close correspondence between the ROI and the area occupied by the tissue.

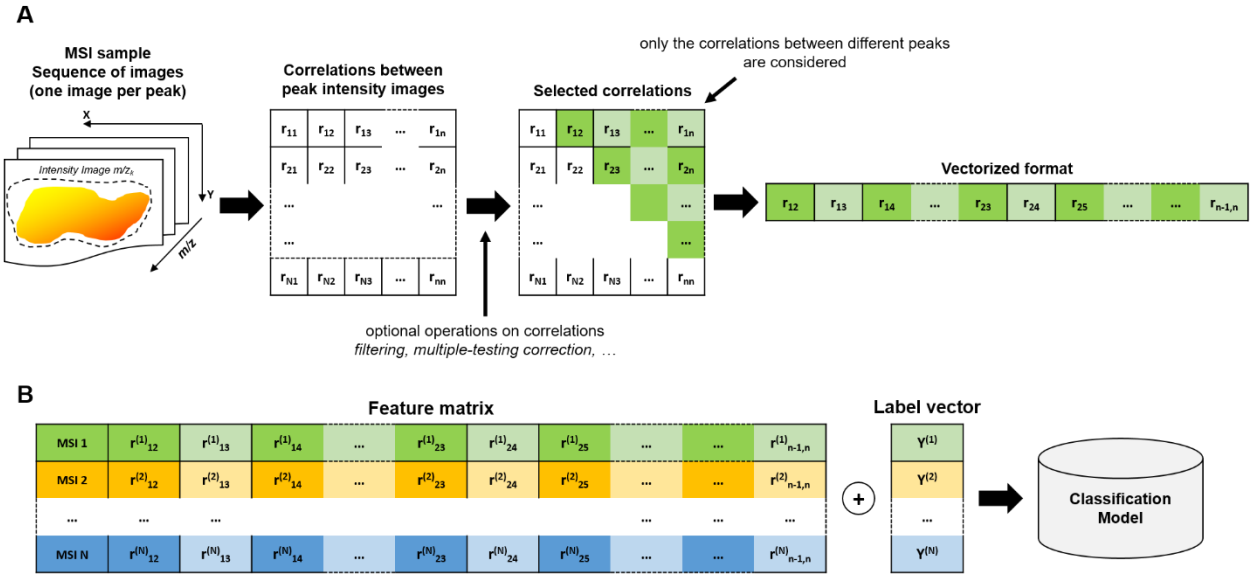

Figure 2 - Co-localization based classification scheme. (A) The correlation matrix is calculated from all the pairs of ion images. The upper triangular matrix elements are vectorized to define the feature vector representing the MSI sample. (B) A classification model can be fitted on the co-localization feature vectors and the corresponding MSI sample labels.

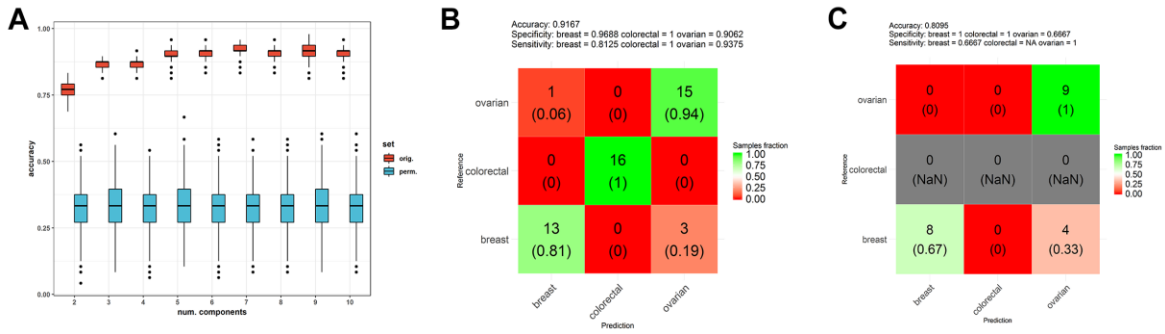

Figure 3 – (A) Boxplot of the PLS-DA accuracies for the 10-fold cross-validation repeated 1,000 times applied to the cross-validation set (red) and compared with the accuracies associated to randomly shuffled labels (cyan). (B) Confusion matrix of a random example of the 1,000 repetitions of the 10-fold cross-validation, using the optimal number of PLS-DA components equal to 7. (C) Confusion matrix of the PLS-DA model applied to the external test set. Although an expected drop in performances is observed, the results show that the model can be used to discriminate between tumor types.

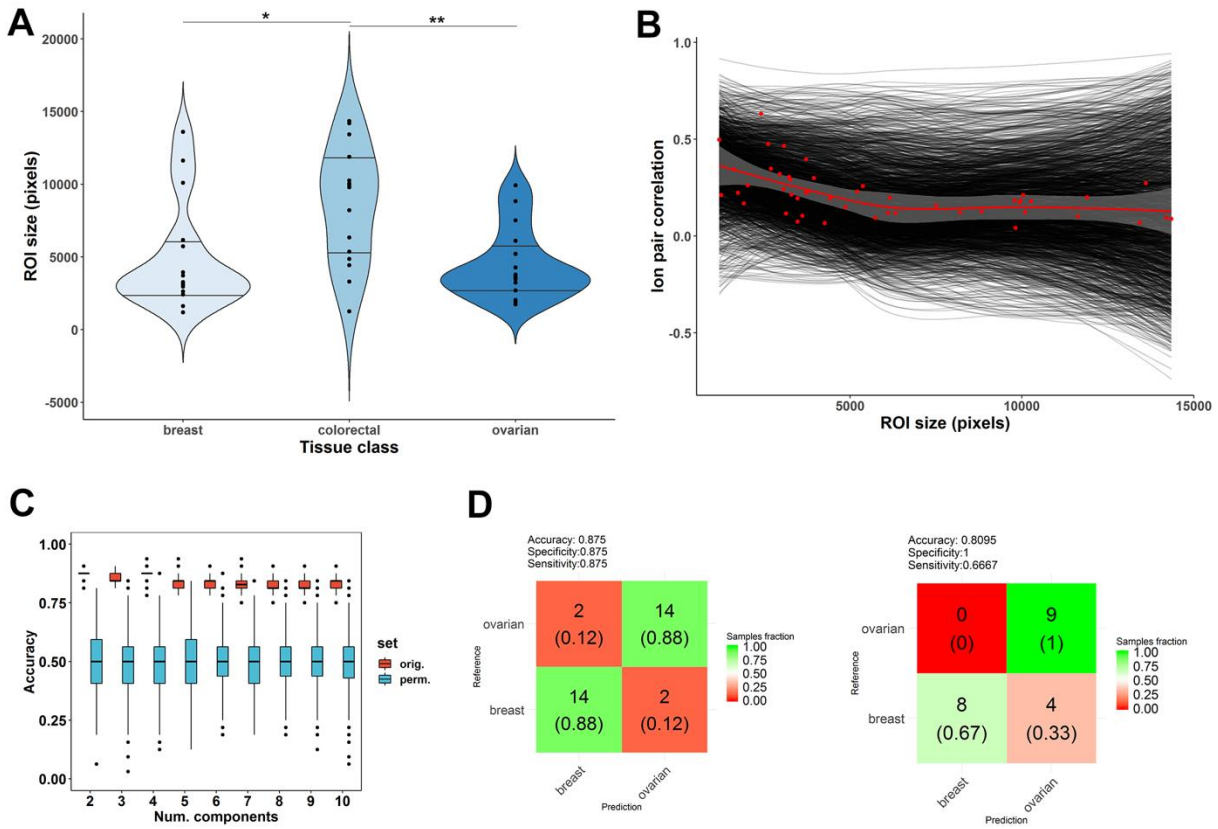

Figure 4 - Analysis of the dependence of correlations and ROI size. The three main tumor classes showed different distributions of ROI sizes (A). The Wilcoxon-Mann-Whitney test rejected the null hypothesis of equal distribution for the pairs 'breast/colorectal' and 'colorectal/ovarian' ($p = 0.013$, $p = 0.008$, respectively), whereas the null hypothesis could not be rejected for the pair 'breast/ovarian'. The LOESS models (B) fitted on individual correlation features versus the ROI size in pixels show a slightly decreasing trend with an elbow point around 5,000 pixels. The red line represents the LOESS model fitted on the average correlations within each image and the red dots the observed average values. The 1,000 repeated 10-fold cross-validations of the breast and ovarian cancer samples resulted in an optimal number of PLS-DA components equal to 2. Additionally, the performances of the models were significantly higher than after shuffling the sample labels ($p < 0.001$) (C). Both 10-fold cross-validation and external test set prediction performance shows that the model is capable of discriminating these two tumor types. These results confirm that the predictive capacity of the models is not biased by the different ROI size of the three tumor types (D).

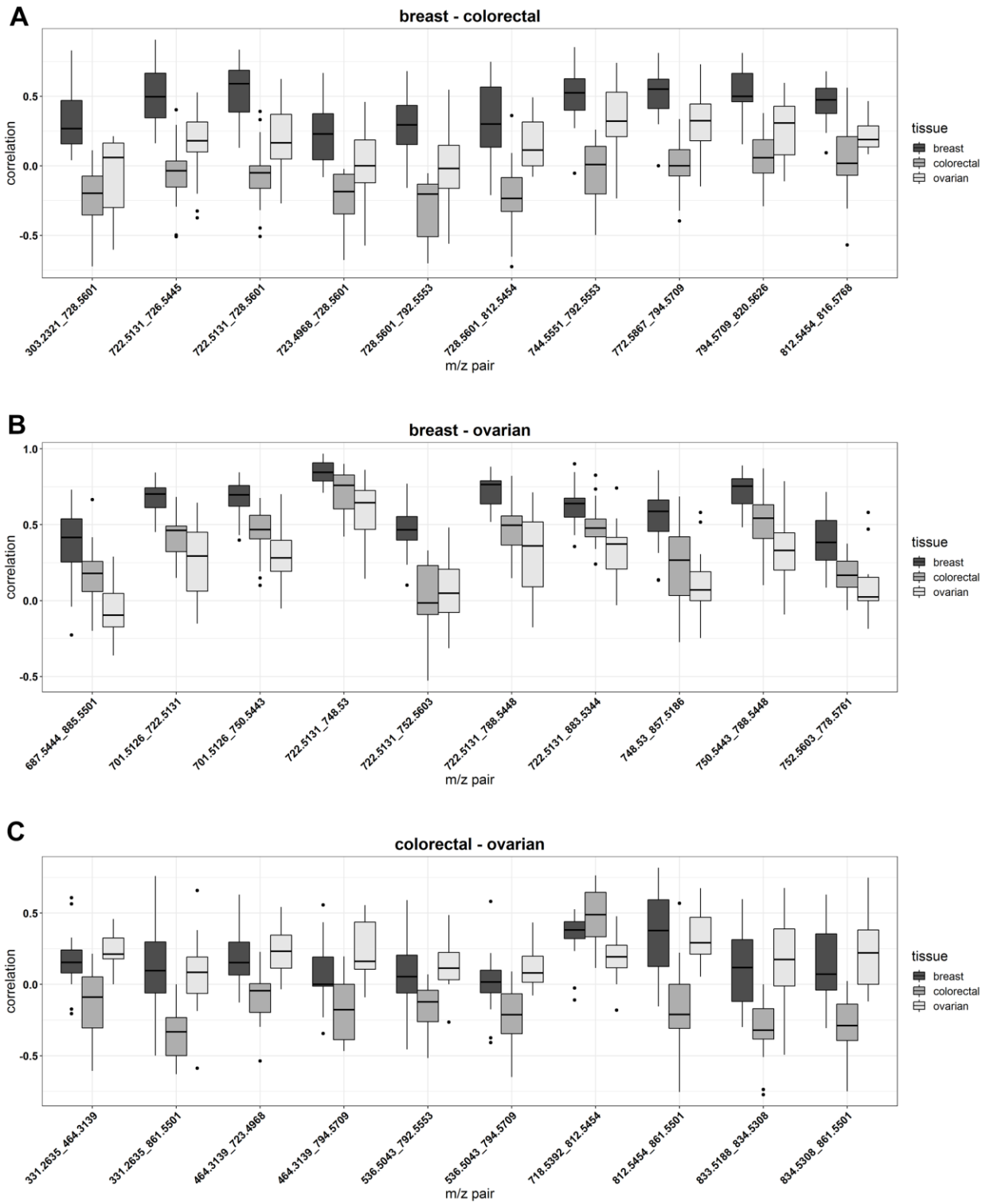

Figure 5 - Boxplot of the ten most significant ion-pairs correlations between breast and colorectal cancer (A), breast and ovarian cancer (B), and colorectal and ovarian cancer (C). The x-axis labels represent the m/z values of the ion pairs, separated by the underscore symbol. All the correlations were significant at a 95% level after Benjamini-Hochberg correction.

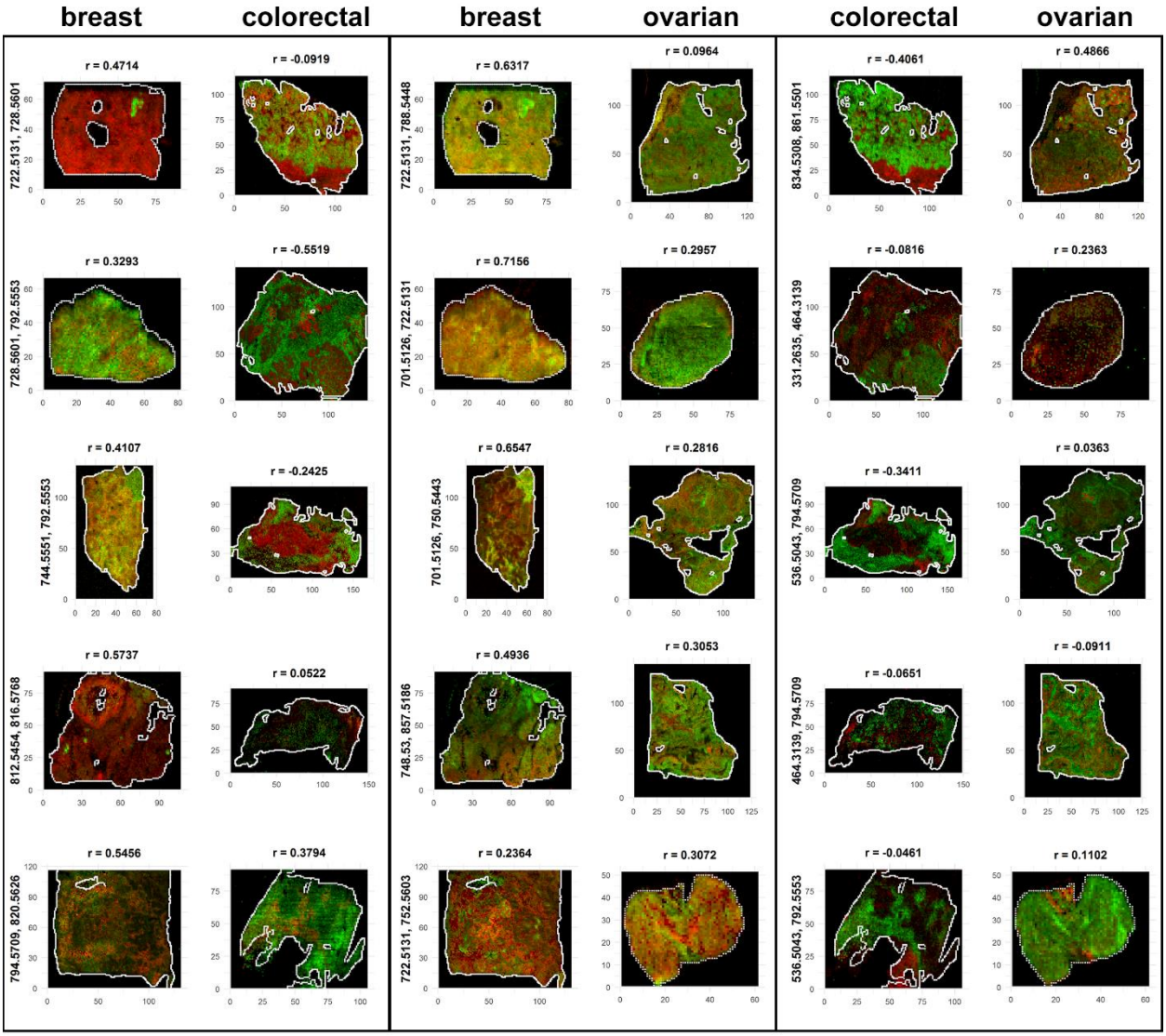

Figure 6 – Comparison of the intensity images of the ions associated with the five most significant correlations for the three pairwise comparisons. The ion intensities are scaled in [0, 1] and plotted as red and green channels, respectively. The reported tissue sections have the closest value of the selected correlation to the median value within the same tumor type subset. The values of the Spearman's correlations are reported on top of each image. The tissue-related ROIs are delineated with a white line.

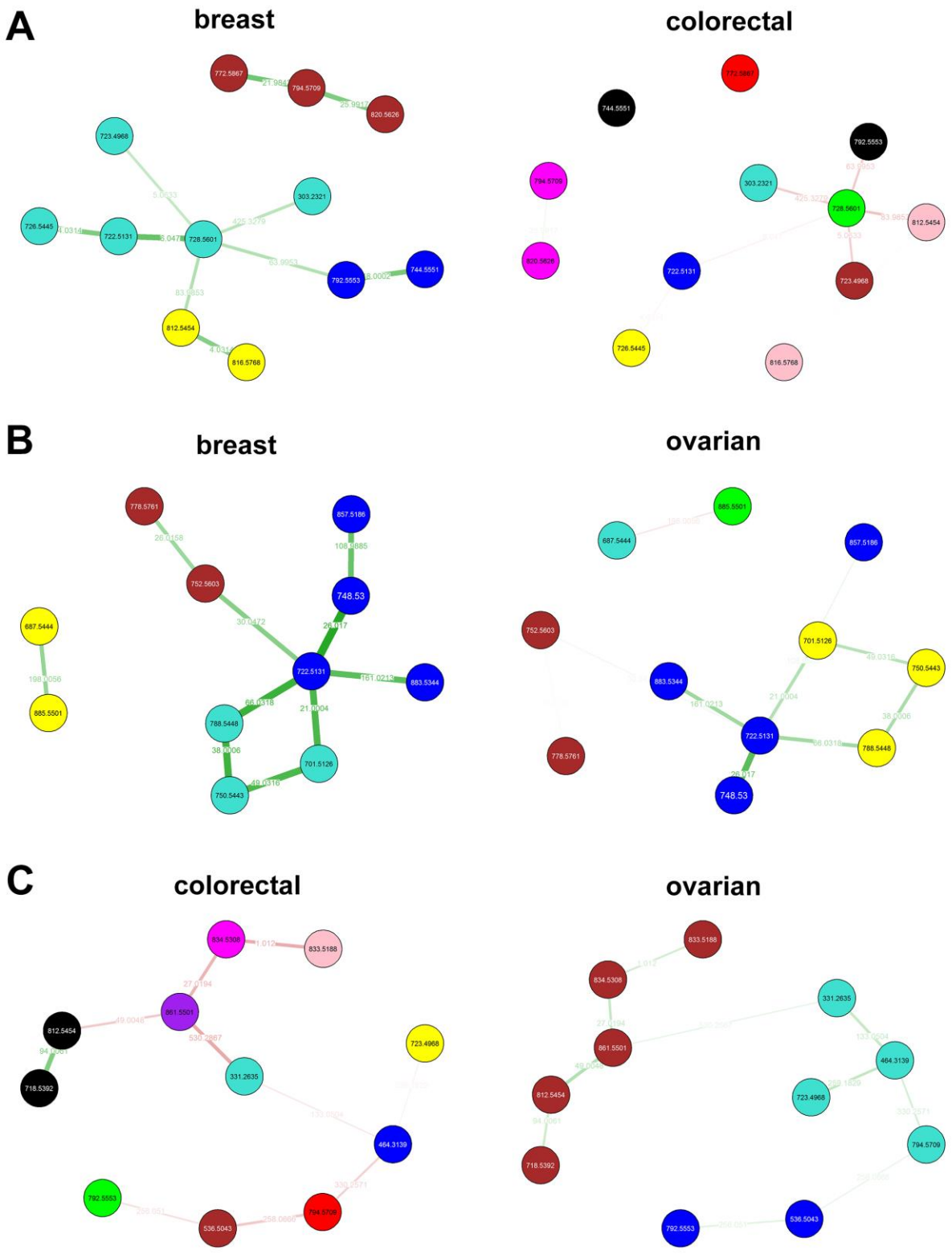

Figure 7 - Weighted graph of the top 10 correlation features in the three pairwise tumor comparisons (A, B, C). In each graph, edges represent the median correlation value across the samples of a tumor type (green = positive, red = negative). Nodes represent the m/z values associated with the correlation

features. The labels on the edges represent the absolute difference in m/z between the nodes m/z values. Community structures of nodes determined by random walk are plotted with the same colors and can be interpreted as members of possible biochemical mechanisms.