Network analysis of mass spectrometry imaging data from colorectal cancer identifies key metabolites common to metastatic development.

Paolo Inglese[1], Nicole Strittmatter[1], Luisa Doria[1], Anna Mroz[1], Abigail Speller[1], Liam Poynter[1], Andreas Danhorn[1], Hiromi Kudo[2], Reza Mirnezami[1], Robert D Goldin[3], Jeremy K Nicholson[1], Zoltan Takats[1†], Robert C Glen[1,4†]

[1] Department of Surgery and Cancer, Division of Integrative Systems Medicine and Digestive Diseases, Imperial College London, SW7 2AZ London (UK)

[2] Department of Medicine, Imperial College London, SW7 2AZ London (UK)

[3] Centre for Pathology, Imperial College, W2 1NY London (UK)

[4] Centre for Molecular Informatics, Department of Chemistry, University of Cambridge, CB21EW Cambridge (UK)

†Contributed equally, correspondence to be addressed to z.takats@imperial.ac.uk and r.glen@imperial.ac.uk

Abstract

A deeper understanding of inter-tumor and intra-tumor heterogeneity is a critical factor for the advancement of next generation strategies against cancer. The heterogeneous morphology exhibited by solid tumors is mirrored by their metabolic heterogeneity. Defining the basic biological mechanisms that underlie tumor cell variability will be fundamental to the development of personalized cancer treatments. Variability in the molecular signatures found in local regions of cancer tissues can be captured through an untargeted analysis of their metabolic constituents. Here we demonstrate that DESI mass spectrometry imaging (MSI) combined with network analysis can provide detailed insight into the metabolic heterogeneity of colorectal cancer (CRC). We show that network modules capture signatures which differentiate tumor metabolism in the core and in the surrounding region. Moreover, module preservation analysis of network modules between patients with and without metastatic recurrence explains the inter-subject metabolic differences associated with diverse clinical outcomes such as metastatic recurrence.

Significance

Network analysis of DESI-MSI data from CRC human tissue reveals clinically relevant co-expression ion patterns associated with metastatic susceptibility. This delineates a more complex picture of tumor heterogeneity than conventional hard segmentation algorithms. Using tissue sections from central regions and at a distance from the tumor center, ion co-expression patterns reveal common features among patients who developed metastases (up of > 5 years) not preserved in patients who did not develop metastases. This offers insight into the nature of the complex molecular interactions associated with cancer recurrence. Presently, predicting CRC relapse is challenging, and histopathologically like-for-like cancers frequently manifest widely varying metastatic tendencies. Thus, the methodology introduced here more robustly defines the risk of metastases based on tumor biochemical heterogeneity.

Author contributions

P.I., Z.T., R.C.G.: designed the study, developed the workflow, analyzed the data, interpreted the results, wrote the paper; N.S. collected the MS, performed the H&E staining, wrote the paper; L.D.: interpreted the results, wrote the paper; A.M.: collected the MS; A.S.: histological assessment; L.P.: collected the tissue specimens and clinical metadata; A.D.: collected the MS; H.K.: performed the H&E staining; R.M.: collected the tissue specimens and clinical metadata. R.G.: histological assessment; J.K.N: designed the study, edited the paper.

Introduction

Cancerous tissue is characterized by a high level of heterogeneity when compared to the surrounding host tissue. This heterogeneity, expressed at multiple levels, may be purely genetic (1) or epigenetic (2) or a combination, and represents one of the most challenging aspects for the implementation of effective anti-cancer treatment strategies. In earlier work, mass spectrometry imaging (MSI) techniques have shown promise in their ability to capture the diversity of tumors through the analysis of molecular expression patterns (3-8), showing that the underlying heterogeneity is reflected in dramatic metabolic changes across different tumor sub-regions. From the point of view of statistical modelling, the lack of detailed metabolic data of the heterogeneous properties of cancerous tissues and the difficulty of detecting those differences through visual inspection of the specimens makes

necessary the use of analytical methods combined with unsupervised analysis techniques. These approaches can detect the underlying statistical structure of the data and can be employed to cluster molecular abundance patterns that are found to be common among discrete regions of tissues. These subsets can be defined both in the spatial domain (pixels) and in the metabolic domain (ions corresponding to molecules) and both approaches have been explored in previous work for the study of cancer and other types of tissues. Linear methods, such as PCA or MDS, combined with hierarchical clustering have shown that MALDI-MSI can capture the heterogeneity of cancer while allowing analysis of the properties of the metabolic space (9, 10), whereas other approaches based on the combination of multiple clustering methods base their results on the presence of concordance between the spatial patterns associated with clusters (pixels) (11, 12). Other studies have associated the presence of specific sets of molecules in the cancerous tissue with clinical outcome (e.g. increased risk of mortality or metastasis development) (13, 14). However, the conclusions from this earlier work are often limited to providing a series of possible biomarkers for tumor heterogeneity - limited or no insight is given into the possible biochemical mechanisms related to the increased/decreased abundance of identified markers in a specific region of the tissue.

We hypothesize that the mechanisms behind tumor heterogeneity are much more complex than those captured simply by the identification of sets of locally highly abundant metabolites and, for this reason, it is extremely important to introduce methodologies capable of capturing a broader view of the underlying biochemical interactions. Network analysis represents a natural choice in this regard, since it provides a set of tools to describe the possible interactions (co-expressions) between variables in terms of their mutual statistical similarities. In the present study, we have characterized the metabolic differences between patients with or without metastatic recurrence using weighted co-expression network analysis (WGCNA) (15) module preservation (16). Previous studies have applied a WGCNA approach to metabolic data (17-19) and, specifically, several studies have made use of module preservation as a measure for the identification of (gene) expression differences between species or disease conditions (20-26) but, to the best of our knowledge, this is the first application to MSI data derived from cancerous tissue specimens. A related study uses correlations to evaluate the ion co-localization in MSI data (27). The hypothesis underlying the present study is that network modules can represent different aspects of the local metabolism through the analysis of co-localized ions and that the metabolic differences characterizing tumor aggressiveness can be identified by those modules associated with patients with metastases that are (or are not) preserved in patients without metastases.

For both the data collected at the tumor core and at a distance of 10cm from the tumor, two groups of patients are defined: a *control cohort* consisting of patients with metastatic recurrence and a *test cohort* consisting of those patients without metastatic recurrence. Using WGCNA, an ion network can be constructed using the mass spectral data from instances where the variables have stronger connections if co-localized (correlated). However, in contrast to gene expression analysis, where each sample can be treated individually, MSI data from a single patient specimen often consists of thousands of mass spectral profiles (one for each pixel). This, together with the different number of patients belonging to the two cohorts, may result in a significantly different number of spectra associated with the two clinical outcomes. Furthermore, treating the spectra as independent measures does not guarantee that the correlations reflect the co-expression levels of the ions within each individual tissue specimen.

For these reasons, a comparative analysis is performed using a consensus cohort network, defined by merging the tissue sample networks from a single cohort. A set of modules are defined by partitioning the consensus network into sub-networks of highly correlated nodes and the metabolic differences between the two cohorts are characterized in terms of module preservation.

In this study, the WGCNA module preservation analysis on DESI-MSI data from specimens surgically removed at the tumor center and at a distance of 10cm from the tumor center (Table 1) shows, using no prior information about the nature of the local tissue (tumor, surrounding healthy tissue or background), that the metastasis associated metabolic patterns within the tumor core microenvironment involve not only the cancerous tissue, but are extended to the neighboring stromal tissue and furthermore, that metastasis associated differences (albeit different ones) in the cellular metabolism are also detected in regions relatively distant from the tumor.

Results

Re-calibration of peak positions, visual assessment of sample quality and ion selection.

The peak lists (centrode data), extracted from the RAW data using the ProteoWizard software (28) were re-calibrated in order to reduce the peak shifts arising from specific instrumental conditions during the MS acquisition. Since a dedicated reference compound for lock-mass correction was not infused during the MS measurement process, a re-calibration procedure was introduced exploiting the characteristics of the analyzed data. Two ions were identified as reference candidates because of their standard abundance across all the samples: palmitate ion, corresponding to 255.2330 *m/z* ($[M-H]^-$), was found homogeneously occurring in the entire layer containing both the tissue and the background; whereas phosphatidylinositol, PI(38:4), corresponding to 885.5499 *m/z* ($[M-H]^-$) was mainly found in the region occupied by the tissue. Using these two reference ions, the *m/z* shift across the entire sample could be estimated. Setting a 1 Δppm wide search window, the intensity of the closest peak to the reference ions was associated with each pixel; if no peak was found within the search window, the pixel was left blank. Because of the characteristics of the two reference ions, it was expected that the intensity of the image corresponding to palmitic acid would cover the entire layer whereas the image corresponding to PI(38:4) would clearly represent the tissue sample. If the value of Δppm was too small, the matched ions intensity image would instead present holes or scattered pixels. In this way, after scanning increasing values of Δppm in steps of 1ppm, the optimal Δppm was defined as the minimum value to produce images defined as having no more than 1% of missing pixels for the palmitic acid image and a clear tissue image for PI(38:4) (Fig. 1A-F). If the optimal Δppm was larger than 10*ppm* (more than double the instrumental error), then the quality of the sample was considered insufficient and the sample was discarded. *S*ubsequently, the *m/z* values found in the search window corresponding to the palmitate peak were used to quantify the peak shift across all the pixels. A robust Locally Weighted Scatterplot Smoothing (LOWESS) model was used to estimate the relative distance between the RAW peak positions and the expected *m/z* value (255.2330 *m/z*) (Fig. 1G). Only the palmitic acid was used to estimate the peak shifts; for this ion, the LOWESS model was also capable of estimating the peak shift in the presence of a small number of blank pixels, whereas it failed to estimate the peak shift for PI(38:4) because of the large number of consecutive blank pixels (Fig. 1H). The estimated peak shift was used to re-calibrate all the spectra, applying a rigid translation of the *m/z* values, so that the *m/z* values corresponding to palmitic acid were set equal to the theoretical value. While this approach gave satisfactory results for Orbitrap data, the same algorithm cannot be directly used for the correction of time-of-flight data.

After peak matching (SI Appendix, Materials and Methods), in order to compare the different co-occurrence patterns, only the ions present in all the samples in at least 1% of the pixels were used. Additionally, in order to consider only the biological correlation patterns, isotopes were identified and removed. An ion at an *m m/z* position was denoted as an isotope if there existed another ion at m_0 in the matched ions list such that $m \in [m_0 - 1.003\,k, m_0 + 1.0045\,k]$, with $k \in \mathbb{N}$ and such that it could not be annotated using an error threshold of 5 ppm were checked.

Modelling the biochemical differences between groups of patients through the non-preserved network ion modules.

In all experimental analyses, two patients' cohorts were defined: the *reference cohort*, X_{met}, consisting of those patients that showed clinical outcomes of interest (developed metastasis), and the *test cohort*, $X_{non\text{-}met}$, consisting of those patients who did not develop metastasis during the follow-up period. The assumption behind this procedure was that the differences in ion abundances in local regions of the tissue reflected metabolic heterogeneity and that differences in metabolic pathways occurring in the cancer cells were associated with a different clinical outcome. In order to calculate a representative network for each cohort, the definition of a consensus network was used (15). A signed adjacency (SI Appendix, Materials and Methods) was calculated from each tissue specimen of the cohort, and combined into a single matrix whose elements were defined as the corresponding minimum value across all the cohort samples. The soft power β, necessary to calculate the signed topological overlap matrix (TOM) was determined as the smallest integer corresponding to a $R^2 \geq 0.8$ for the scale-free network assumption for both X_{met} and $X_{non\text{-}met}$ in a range of 1 to 20. Using the same value of β aimed to preserve the compatibility between the control and test networks. In a similar way, the consensus TOM from the only control data, denoted TOM_{met}, was defined as the matrix whose elements were equal to the corresponding minimum values across the *TOM*s of the cohort (SI Appendix, Materials and Methods).

In order to obtain a partition of disjoint sets of highly correlated ions (denoted *reference modules*) in the reference network, a hierarchical clustering with average linkage on the TOM_{met} based distance matrix was applied. The distance matrix, calculated as 1-TOM_{met}, represented the dissimilarities between the ion expressions within the metastatic tissue specimens. The optimal partition was identified using the Dynamic Tree Cut hybrid algorithm (29) with a minimum number of ions per cluster equal to 5. The latter choice was based on the hypothesis that small clusters could capture a more detailed picture of the spatial similarities in the molecular expressions. The module eigenmetabolites (ME) (in a similar fashion to the module eigengenes of the original WGCNA algorithm, these were defined as the scores of the first principal component of the MSI data limited to the module ions) defined the representative spatial distribution of the ions belonging to the module. In order to reduce the potential redundancy of similar ion modules, those modules where the MEs were characterized by a Pearson's correlation larger than 0.8 were merged into a single module. The ions that were not clustered (because they were assigned to modules with less than 5 ions) were assigned to the "grey" module and were not passed to the next stage of the analysis. In this way, a set of modules were associated to the consensus network, representing the observed molecular heterogeneity of the cancerous tissue. The metabolic differences between the reference cohort of patients and test cohort of patients were determined by module preservation analysis (SI Appendix, Materials and Methods). The X_{met} consensus network modules were tested for preservation in the $X_{non\text{-}met}$ consensus network, under the hypothesis that not-preserved modules would capture the metabolic differences associated with the patients' prognosis.

Phosphatidylglycerol metabolism in stromal tissue surrounding the tumor core is associated with metastasis.

The mass spectral data from the tumor core specimens of 32 subjects (Table 1) was used to perform the module preservation analysis. The MSI data from the 8 patients that developed metastasis, denoted $X_{met}^{(core)}$, was used to define the consensus reference network, while the remaining 24 patients' MSI data, denoted $X_{non-met}^{(core)}$, was used to build the test network. The high correlation between their node connectivity *k*, defined as the sum of the node edges strengths (30), (Pearson's r = 0.95, Fig. 2A) showed that the two networks were characterized by similar global properties. This result showed that the global tumor metabolism was undifferentiated between the two groups of patients. By partitioning the $TOM_{met}^{(core)}$ using the hierarchical clustering, 6 reference modules were found (denoted as "blue", "brown", "green", "red", "turquoise" and "yellow") (Fig. 2A). A visual inspection of the ME images showed that 4 modules were associated with the tissue and 2 were localized mainly outside of the tissue (Fig. S2). Moreover, it was seen that 3 modules captured different tumor sub-regions and one module was mainly expressed in the stromal connective tissue (Fig. 3A). The combination of the relative intensities of the MEs localized within the tumor region revealed the complex spatial patterns associated with its molecular heterogeneity (Fig 3A). The module preservation analysis (5,000 permutations) showed that 4 of the 6 metastatic network modules were from weakly to moderately preserved in the $X_{non-met}$ network. Among those, the green module was associated with tissue and was weakly preserved ($Z_{summary} \lesssim 5$) in the $X_{non-met}$ data (Fig. S1A). The relatively lower correlation (Pearson's r = 0.65) between the $X_{met}^{(core)}$ and $X_{non-met}^{(core)}$ module membership (*kME*), defined as the correlation between the ions intensities and the ME, confirmed that the green module ions were not equally co-localized in the two datasets (Fig. S1A). The projection of its ME intensities on the optical image of the H&E stained tissue showed that the green module ions involved the stromal tissue segment of the tumor with the observed presence of free fat droplets and infiltrations of tumor cells (an example is reported in Fig 4A). Noticeably, the MEs revealed that the green module ions were also expressed in the $X_{non-met}^{(core)}$ samples (Fig. S3) confirming that the relevant biochemical differences could be captured only in terms of different molecular interactions (network edge weights) and not in terms of presence/absence of certain sets of ions. Analyzing the corresponding regions in the optical images of the H&E stained tissue, it was observed that this module corresponded to regions close to the necrotic tissue (an example is reported in Fig. 4B).

The ion annotation, performed through a Metlin database search (31), revealed that those of the weakly preserved green module consisted of phosphatidylglycerols (PG) and fatty acids (FA) (Fig 5A, Table S1). Specifically, hexadecanoic acid (C16:0), eicosatetraenoic acid (C20:4), docosatetraenoic acid (C22:4), docosapentaenoic acid (C22:5), docosahexaenoic acid (C22:6) and tetracosapentaenoic acid (C24:5) which showed high correlation with the distribution of PG(36:3), PG(36:4), PG(38:5), PG(38:6), PG(40:6), PG(40:7) and the non-annotated 292.2432, 528.2736 and 791.5441 *m/z* ions. The ion pair specificity (SI Appendix, Materials and Methods), confirmed that most of the PGs together with tetracosapentaenoic acid were specific of the $X_{met}^{(core)}$ sample set (Fig. 5A).

Confirmation of the different network properties between the discovery and test datasets was obtained by running WGCNA directly on the $X_{non-met}^{(core)}$ samples. Indeed, it was seen that the green module ions were split into two different non-metastatic modules (Fig. S4C). Most of the FAs were assigned to a different module together with phosphatidylethanolamines (PEs), phosphatidylinositols (PIs) and phosphoserines (PSs) (Table S2, Fig. S4D). The PGs were instead assigned together in a larger module consisting of only phospholipids (Table S3, Fig. S4E).

In order to validate and compare the statistical relationships between the module network topologies and the classification of "metastatic" vs "non-metastatic" tissue sections, a leave-one-specimen-out classification was performed, using the intra modular connectivity as feature vectors (SI Appendix, Materials and Methods). A Regularized Random Forest (RRF) classifier (32) trained on the "green" intramodular connectivity was capable of predicting the metastatic tissue sections with a balanced accuracy (33) of 0.8958 (Fisher test p-value < 8e-5) (Table S4). Since the soft power and the modules obtained by the consensus network were used for the calculation of the intramodular connectivity, this test was performed with the sole objective of confirming that the consensus network was capable of capturing the local properties of single patients' metabolism, and these results do not represent the predictive power on unseen data.

Phospholipid metabolism is different at 10cm from the tumor core of metastatic and non-metastatic patients.

Analogous to the tumor core analysis, module preservation analysis was performed using the MSI data from tissue samples collected from surgical specimens at 10cm from the tumor core. The specimens from 29 patients (Table 1) were split into two groups corresponding to the occurrence/non-occurrence of a metastatic relapse during the follow-up period. The MSI data of the 6 patients who developed metastases, denoted $X_{met}^{(10cm)}$, were used to determine the reference network modules that were later tested for preservation in the remaining 23 patients MSI data ($X_{non-met}^{(10cm)}$). As seen in the tumor center samples, the WGCNA consensus networks calculated from the mass spectral data from $X_{met}^{(10cm)}$ and $X_{non-met}^{(10cm)}$ were characterized by similar global properties (Pearson's r = 0.93 between metastatic and non-metastatic data *k*, Fig. 2B); this result also showed that in this case the two metabolic networks were globally almost identical. The metastatic related network was clustered in 6 modules (denoted "blue", "brown", "green", "turquoise", and "yellow") (Fig 2B). Analyzing the spatial distribution of the detected ME, it was observed that three of those modules showed were localized in the tissue of all 6 patients (Fig. S5). The combination of the relative intensity of those MEs revealed the molecular heterogeneity of the tissue sections (Fig. 3B). Using module preservation analysis with 5,000 permutations, it was observed that all the modules were from weakly to moderately preserved ($2 \leq Z_{summary} \leq 10$). In particular, the brown module, expressed in the tissue, was weakly preserved ($Z_{summary} \lesssim 5$) in the patients who did not develop metastasis. As observed in the tumor core data, the $X_{met}^{(10cm)}$ and $X_{non-met}^{(10cm)}$ the brown module *kMEs* showed a relatively lower correlation (Pearson's r = 0.74, Fig. S1B).

The annotation of the brown module ions, mainly expressed in the epithelium (as revealed by a comparison with the optical images of the H&E stained images, shown in Fig. 4C-D), consisted of different classes of phospholipids (PE, PG, PI, and PS) together with phosphatidic acids (PA), octadecadienoic acid (C18:2), eicosenoic acid (C20:1), eicosadienoic acid (C20:2), and eicosatetraenoic acid (C20:4) (Table S5). The ion pair significance showed that the phospholipid distributions showed stronger correlations in $X_{met}^{(10cm)}$ than in $X_{non-met}^{(10cm)}$ samples (Fig 5B), with PG(36:4) playing a central role in the metastatic network specific edges.

Running WGCNA directly on the $X_{non-met}^{(10cm)}$ MSI data showed that most of the brown module ions were assigned to a more complex module, labelled "turquoise", together with different classes of phospholipids and FAs (Fig. S7A-C). Specifically, among the annotated ions, the most probable candidates for 281.2486, 279.233, 277.2174, 309.28, 306.2521 *m/z* were identified to be n-9 oleic acid (C18:1), n-6 linoleic acid (C18:2), n-3/6 α/γ-linoleic acid (C18:3), n-9 eicosenoic acid (C20:1), n-6

273 dihomo-γ-linoleic acid (C20:3) and n-3 docosahexaenoic acid (C22:6); those ions were found co-
274 localised with the phospholipids (Table S6, Fig. S7D).

275 As observed in the tumor core analysis, those ions were also expressed in the test samples (Fig. S6),
276 confirming the importance of the preservation analysis in order to determine the differences between
277 the metastatic and the non-metastatic samples in terms of local metabolite interactions.

278 In a similar fashion to previous observations in the tumor core analysis (the RRF classifier trained on
279 intramodular connectivity), it was seen that the weakly preserved brown module provided the best
280 performance in discriminating between "metastatic" and "non-metastatic" tissue sections with a
281 balanced accuracy of 0.8080 (Fisher test p-value = 0.0105) (Table S7), confirming that the consensus
282 network was capable of capturing the metabolic characteristics of single patients.

283 Discussion

284 Understanding the mechanisms behind the heterogeneity of cancerous tissues is a key goal for the
285 development of effective cancer treatment strategies (34). In this study, a network based analysis of
286 the co-localized ions allowed the identification of a set of tissue sub-types (ion modules) that
287 represent the tissue heterogeneity associated with each individual subject. The WGCNA approach,
288 combined with the Dynamic Tree Cut algorithm, determined the optimal number of tissue sub-types
289 with no user intervention. The only parameter required, the minimum number of ions necessary to
290 form a module, was set to 5 following the hypothesis that small clusters could capture finer details of
291 the heterogeneous metabolic patterns. In order to detect the metabolic differences between the
292 group-of-interest (patients developing metastasis) and the remaining part of the cohort, the modules
293 detected were tested in the former for preservation in the latter group. The permutation test applied
294 on MSI data at a distance of 10cm from the tumor core allowed the identification of one module
295 associated with the development of metastasis in both the cases.

296 A set of molecular interactions involving PG and FAs was found to be representative of the increased
297 risk of the development of metastases in the tumor core. Specifically, the combination of WGCNA and
298 the differential network analysis showed that, in the metastatic tumor core set, PGs were more highly
299 co-localized together with tetracosapentaenoic acid than in the non-metastatic related samples,
300 suggesting a key role for these molecules in differentiating the local biochemical reactions. The
301 presence of polyunsaturated fatty acids (PUFAs) in the "brown" module together with PGs suggests
302 that these regions elicit significantly higher phospholipase A2 (PLA2) activity. This enables the
303 formation of prostaglandins, in particular prostaglandin E2, through COX2 activity, which has widely
304 been associated with tumor invasion and metastasis formation in a broad range of different cancers.
305 cPLA2 activity itself has been found to contribute to metastasis formation in breast cancer *via* TGF β-
306 induced epithelial-mesenchymal transition (EMT) (35).

307 PGs are generally considered to be membrane constituents of bacterial cells, serving mainly as
308 surfactants in mammalian organisms. In this regard, the PGs detected are unlikely to be building blocks
309 of cellular membranes in the tumor environment, but rather present in the interstitial space.
310 Considering the apparent specificity of PLA2 for anionic phospholipids likely excreted into the
311 extracellular space and the rich abundance of macrophages in association with the green module (Fig.
312 4A), the underlying enzyme is probably the PLA2 produced by macrophages. Macrophage cPLA2 has
313 already been associated with metastasis formation in case of lung cancer and it was successfully
314 demonstrated that depletion of macrophage cPLA2 by gene knock out resulted in the significant
315 decrease of metastatic potential (36). In conclusion, the macrophage infiltration clearly shown on Fig.

316 4A is likely to confer sufficient PLA2 activity to trigger EMT and downstream metastasis formation in
317 the corresponding group. Although the involvement of gut microbiota is not obvious in the current
318 case, bacterial PGs, as the main constituents of prokaryotic cellular membranes can potentially
319 contribute to the expression of macrophageal PLA2 through TLR2 pathway, especially in the proximity
320 of necrotic areas.

321 From analysis of the weakly preserved module between the metastatic and non-metastatic related
322 tissue samples at 10cm from the tumor core, it was found that the PUFAs octadecadienoic acid,
323 eicosenoic acid, eicosadienoic acid and eicosatetraenoic acid were correlated mainly with PG, PE and
324 plasmalogen PE together with PS and PI in sub-regions of epithelial tissue. In this case, the significantly
325 stronger co-expression levels of the phospholipids determined by their high significance in the
326 metastatic related tissues can be interpreted as the result of increased activity of a number of enzymes
327 including phospholipase D (PLD). Abnormalities in PLD expression can be responsible for altered cell
328 proliferation mechanisms (37, 38). Similarly to the tumor core, a relatively high correlation of
329 eicosatetraenoic acid with most of the detected phospholipids was observed in the weakly preserved
330 module, suggesting evidence of PLA2 activity that can be interpreted as a localized inflammatory
331 condition, also confirmed by the visual inspection of the corresponding regions in the H&E stained
332 tissues (Fig. 4B). This, combined with the abundance of plasmalogens in the module, implies
333 peroxisomal involvement in metastasis formation. It has recently been demonstrated that
334 peroxisomal functionality is key for macrophage activation (39) which can explain the correlation
335 between the peroxisomal metabolic phenotype represented by the network structures (Fig 5B) and
336 metastatic potential conferred by macrophage PLA2 activity.

337 Conclusions

338 Tumors are characterized by a wide range of metabolic, genetic, and phenotypic properties. This
339 heterogeneity, expressed at the genetic and epigenetic level, is reflected by the metabolites produced
340 within cells. In this study, DESI-MSI was shown to be effective for untargeted analysis of the local
341 metabolism of colorectal cancer tissue. The differences in the correlations between the distribution
342 of key ions and their preservation in the network modules identified that the increased activity of
343 PLA2, exploiting the high availability of PG in the surfactant layer in the tumor center, was associated
344 with a risk of post-surgical distant metastasis formation. The molecular correlations which were not
345 preserved revealed that eicosanoid precursors in stromal tissue can play a significant role in the
346 increased cellular proliferation within the tumor core. At 10cm from the tumor center, the altered
347 peroxisomal metabolic phenotype was found to be associated with the different metastatic potential,
348 likely *via* a macrophage activation mechanism. Finally, the network approach identified a complex
349 picture of tumor heterogeneity, confirming that a simple 'presence/absence of ions' approach is not
350 able to determine the differences between the metabolic pathways.

351 Due to the limited number of patients considered, these results should be considered preliminary and
352 await confirmation in a larger patient cohort. In future work, the study will be expanded to a larger
353 number of patients, allowing a more detailed analysis of the metabolic differences associated with a
354 range of clinical characteristics. Furthermore, MSI data in combination with local genetic expression
355 identified from spatial ME maps will be included in order to investigate both genetic and metabolic
356 changes to achieve deeper insight of the nature of tumor heterogeneity and its chemical/metabolic
357 interactions with the surrounding tissue.

Materials and Methods

DESI imaging data was acquired from sections of individual tumor central regions and from section of tissue at 10cm from the tumor central regions of, respectively, 32 and 29 subjects affected by colorectal cancer (SI Appendix, Materials and Methods). The tissue specimens were collected after the surgical removal of the tumor. The subjects were monitored after surgery for a period of up to 6 years, and events such as disease occurrences (metastases, or local recurrence) and deaths were recorded. During the observational period, 8 of the 32 subjects and 6 of the 29 subjects developed metastatic recurrence. The tissue sections were fresh-frozen and mass spectrometry imaging (MSI) data were acquired using an automated 2D DESI source mounted to a Thermo Exactive Orbitrap mass spectrometer. MS data were acquired in the negative ion mode, in the range of 200-1,050 *m/z*. The RAW spectra, corresponding to tissue regions (pixels) were pre-processed in order to extract a set of variables associated with the informative peaks and their relative abundances. Centroided data were extracted using the ProteoWizard software (28) and pre-processed using the 'MALDIquant' package for R (40, 41). Before pre-processing, a re-calibration procedure was applied in order to reduce the peak shift across the pixels of each tissue sample. The procedure is described in the 'Results' section. The details of the pre-processing workflow are reported in SI Appendix, Materials and Methods. The tumor core pre-processed data corresponded to a series of matrices (pixels x ions) containing the intensities of the 185 *m/z* values common to all the sample pixels. For the samples at 10cm from the tumor, 141 *m/z* values were found common to all the samples. The consensus adjacency matrix elements were defined as the minimum of the ions pairwise Pearson's correlations associated with each tissue section. In a similar fashion, the TOM was calculated using the signed adjacency raised to the soft power value. Modules were identified by applying the Dynamic Tree Cut hybrid algorithm to the hierarchical dendrogram with an average linkage obtained using 1-*TOM* as distance matrix. Module preservation was performed using the WGCNA '*modulePreservation*' command with 5000 permutations, using the consensus adjacency matrix as input. The module identification procedure was applied directly on the consensus matrices of the non-metastatic related MSI data for both the tumor core and 10cm datasets.

Datasets and scripts are publicly available at https://doi.org/10.4121/uuid:f06dee0d-1d2e-4d67-b978-5bac4087d346.

Acknowledgements

Full ethical approval was obtained from the institutional review board at Imperial College Healthcare NHS Trust (REC reference numbers 07/H0712/112, 11/LO/0686). The project was funded by the European Research Council (ERC Consolidator Grant 'MASSLIP' Contract No: 617896), the National Institute for Health Research through the Imperial Biomedical Research Centre and Cancer Research UK Grand Challenge Program (Project title: "A Complete Cartography Through Multiscale Molecular Imaging").

References

1. Burrell RA, McGranahan N, Bartek J, & Swanton C (2013) The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* 501(7467):338-345.
2. Marusyk A, Almendro V, & Polyak K (2012) Intra-tumour heterogeneity: a looking glass for cancer? *Nat Rev Cancer* 12(5):323-334.

3. Schwamborn K & Caprioli RM (2010) Molecular imaging by mass spectrometry—looking beyond classical histology. *Nature Reviews Cancer* 10(9):639-646.
4. Alexandrov T, Becker M, Guntinas-Lichius O, Ernst G, & Eggeling F (2013) MALDI-imaging segmentation is a powerful tool for spatial functional proteomic analysis of human larynx carcinoma. *J Cancer Res Clin Oncol* 139.
5. Eberlin LS*, et al.* (2010) Cholesterol sulfate imaging in human prostate cancer tissue by desorption electrospray ionization mass spectrometry. *Analytical chemistry* 82(9):3430-3434.
6. Eberlin LS*, et al.* (2012) Classifying Human Brain Tumors by Lipid Imaging with Mass Spectrometry. *Cancer Research* 72(3):645-654.
7. McDonnell LA*, et al.* (2010) Peptide and protein imaging mass spectrometry in cancer research. *Journal of proteomics* 73(10):1921-1944.
8. Schwartz SA, Weil RJ, Johnson MD, Toms SA, & Caprioli RM (2004) Protein profiling in brain tumors using mass spectrometry. *Clinical Cancer Research* 10(3):981-987.
9. Deininger S-O, Ebert MP, Fütterer A, Gerhard M, & Röcken C (2008) MALDI Imaging Combined with Hierarchical Clustering as a New Tool for the Interpretation of Complex Human Cancers. *Journal of Proteome Research* 7(12):5230-5236.
10. Rauser S*, et al.* (2010) Classification of HER2 Receptor Status in Breast Cancer Tissues by MALDI Imaging Mass Spectrometry. *Journal of Proteome Research* 9(4):1854-1863.
11. Jones EA*, et al.* (2011) Multiple Statistical Analysis Techniques Corroborate Intratumor Heterogeneity in Imaging Mass Spectrometry Datasets of Myxofibrosarcoma. *PLoS ONE* 6(9):e24913.
12. Balluff B*, et al.* (2015) De novo discovery of phenotypic intratumour heterogeneity using imaging mass spectrometry. *J Pathol* 235(1):3-13.
13. Abdelmoula WM*, 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):12244-12249.
14. Lou S*, et al.* (2016) High-grade sarcoma diagnosis and prognosis: Biomarker discovery by mass spectrometry imaging. *PROTEOMICS* 16(11-12):1802-1813.
15. Langfelder P & Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics* 9(1):559.
16. Langfelder P, Luo R, Oldham MC, & Horvath S (2011) Is My Network Module Preserved and Reproducible? *PLOS Computational Biology* 7(1):e1001057.
17. DiLeo MV, Strahan GD, den Bakker M, & Hoekenga OA (2011) Weighted correlation network analysis (WGCNA) applied to the tomato fruit metabolome. *PLoS One* 6(10):e26683.
18. Su Y*, et al.* (2014) Metabolomic and network analysis of astaxanthin-producing Haematococcus pluvialis under various stress conditions. *Bioresource technology* 170:522-529.
19. Yu X*, et al.* (2015) Identification and mechanism analysis of chemical modulators enhancing astaxanthin accumulation in Haematococcus pluvialis. *Algal Research* 11:284-293.
20. Miller JA, Horvath S, & Geschwind DH (2010) Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. *Proceedings of the National Academy of Sciences* 107(28):12698-12703.
21. Chen C*, et al.* (2013) Two Gene Co-expression Modules Differentiate Psychotics and Controls. *Molecular psychiatry* 18(12):1308-1314.
22. Oldham MC, Horvath S, & Geschwind DH (2006) Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proceedings of the National Academy of Sciences* 103(47):17973-17978.
23. Xue Z*, et al.* (2013) Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* 500(7464):593-597.
24. Horvath S*, et al.* (2012) Aging effects on DNA methylation modules in human brain and blood tissue. *Genome Biology* 13(10):R97.

25. Huang K, Maruyama T, & Fan G (2014) The Naive State of Human Pluripotent Stem Cells: A Synthesis of Stem Cell and Preimplantation Embryo Transcriptome Analyses. *Cell Stem Cell* 15(4):410-415.
26. Tong M*, et al.* (2013) A Modular Organization of the Human Intestinal Mucosal Microbiota and Its Association with Inflammatory Bowel Disease. *PLOS ONE* 8(11):e80702.
27. McDonnell LA, van Remoortere A, van Zeijl RJM, & Deelder AM (2008) Mass Spectrometry Image Correlation: Quantifying Colocalization. *Journal of Proteome Research* 7(8):3619-3627.
28. Kessner D, Chambers M, Burke R, Agus D, & Mallick P (2008) ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics (Oxford, England)* 24(21):2534-2536.
29. Langfelder P, Zhang B, & Horvath S (2008) Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* 24(5):719-720.
30. Oldham M, Horvath S, & Geschwind D (2006) Conservation and Evolution of Gene Co-expression Networks in Human and Chimpanzee Brains. *Proc Natl Acad Sci USA* 103.
31. Smith CA*, et al.* (2005) METLIN: a metabolite mass spectral database. *Ther Drug Monit* 27:747-751.
32. Deng H & Runger G (2013) Gene selection with guided regularized random forest. *Pattern Recognition* 46(12):3483-3489.
33. Brodersen KH, Ong CS, Stephan KE, & Buhmann JM (2010) The balanced accuracy and its posterior distribution. *Pattern recognition (ICPR), 2010 20th international conference on*, (IEEE), pp 3121-3124.
34. Murugaesu N, Chew SK, & Swanton C (2013) Adapting Clinical Paradigms to the Challenges of Cancer Clonal Evolution. *The American Journal of Pathology* 182(6):1962-1971.
35. Chen L*, et al.* (2017) cPLA2[alpha] mediates TGF-[beta]-induced epithelial-mesenchymal transition in breast cancer through PI3k/Akt signaling. *Cell Death Dis* 8:e2728.
36. Weiser-Evans MCM*, et al.* (2009) Depletion of Cytosolic Phospholipase A₂ in Bone Marrow–Derived Macrophages Protects against Lung Cancer Progression and Metastasis. *Cancer Research* 69(5):1733.
37. Foster DA & Xu L (2003) Phospholipase D in Cell Proliferation and Cancer. *Molecular Cancer Research* 1(11):789-800.
38. Park JB*, et al.* (2012) Phospholipase signalling networks in cancer. *Nat Rev Cancer* 12(11):782-792.
39. Di Cara F, Sheshachalam A, Braverman NE, Rachubinski RA, & Simmonds AJ (Peroxisome-Mediated Metabolism Is Required for Immune Response to Microbial Infection. *Immunity* 47(1):93-106.e107.
40. Gibb S & Strimmer K (2012) MALDIquant: a versatile R package for the analysis of mass spectrometry data. *Bioinformatics* 28(17):2270-2271.
41. Ihaka R & Gentleman R (1996) R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics* 5(3):299-314.
42. Bozza PT & Viola JPB (2010) Lipid droplets in inflammation and cancer. *Prostaglandins, Leukotrienes and Essential Fatty Acids (PLEFA)* 82(4):243-250.

497 # Tables

Tumor center samples	
Age (years):	
median, range	72, [52, 88]
Genre:	
Female	12 (met = 2, (25%))
Male	20 (met = 6, (75%))
Lymph nodes (number):	
median, range	1, [0, 6] met: 2, [0, 6]
Metastasis:	
Bones	1
Liver	6
Lung	3
Omental/Peritoneal	4

10cm samples	
Age (years):	
median, range	72, [55, 87]
Genre:	
Female	11 (met = 1, (17%))
Male	18 (met = 5, (83%))
Lymph nodes (number):	
median, range	0, [0, 5] met: 2, [0, 5]
Metastasis:	
Bones	1
Liver	3
Lung	1
Omental/Peritoneal	4

498

499 *Table 1 – Patient clinical metadata associated with the two datasets.*

500

501

502

503

504

505

Figures

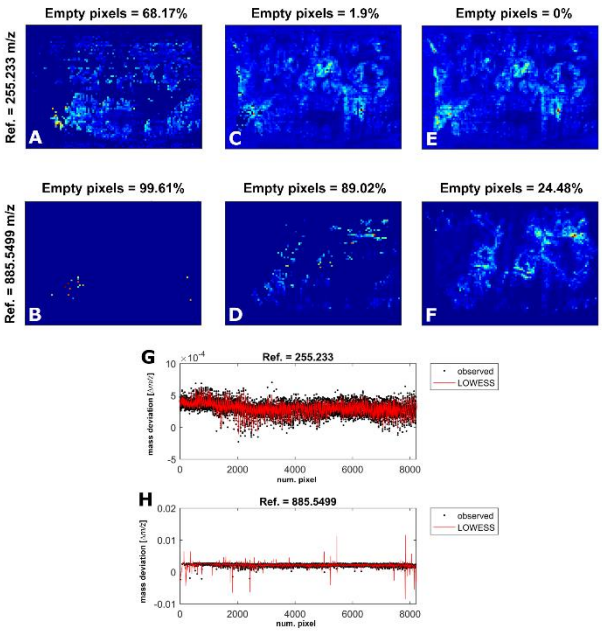

Fig. 1 – Effect of the search window width (Δppm) on the peaks detected in the raw data. In the example shown in the figure, a Δppm equal to 1 or 2 is too small compared to the maximum peak shift found in the sample, producing scattered images for both the palmitic acid and PI(38:4) ($\Delta ppm = 1$, A-B; $\Delta ppm = 2$, C-D). In contrast, using the value $\Delta ppm = 3$, a peak corresponding to 255.2330 m/z for the full layer (0% empty pixels, E) can be detected and produced an image similar to the stained tissue sample (24.48% empty pixels, F). The robust LOWESS model fitted to the peak shift measured with $\Delta ppm = 3$ captures the m/z shift across all the pixels (G), whereas it fails to estimate the m/z shift for the peaks corresponding to 855.5499 m/z because of the large number of missing pixels and also because of the gaps related to the localized presence of tissue in the image (H).

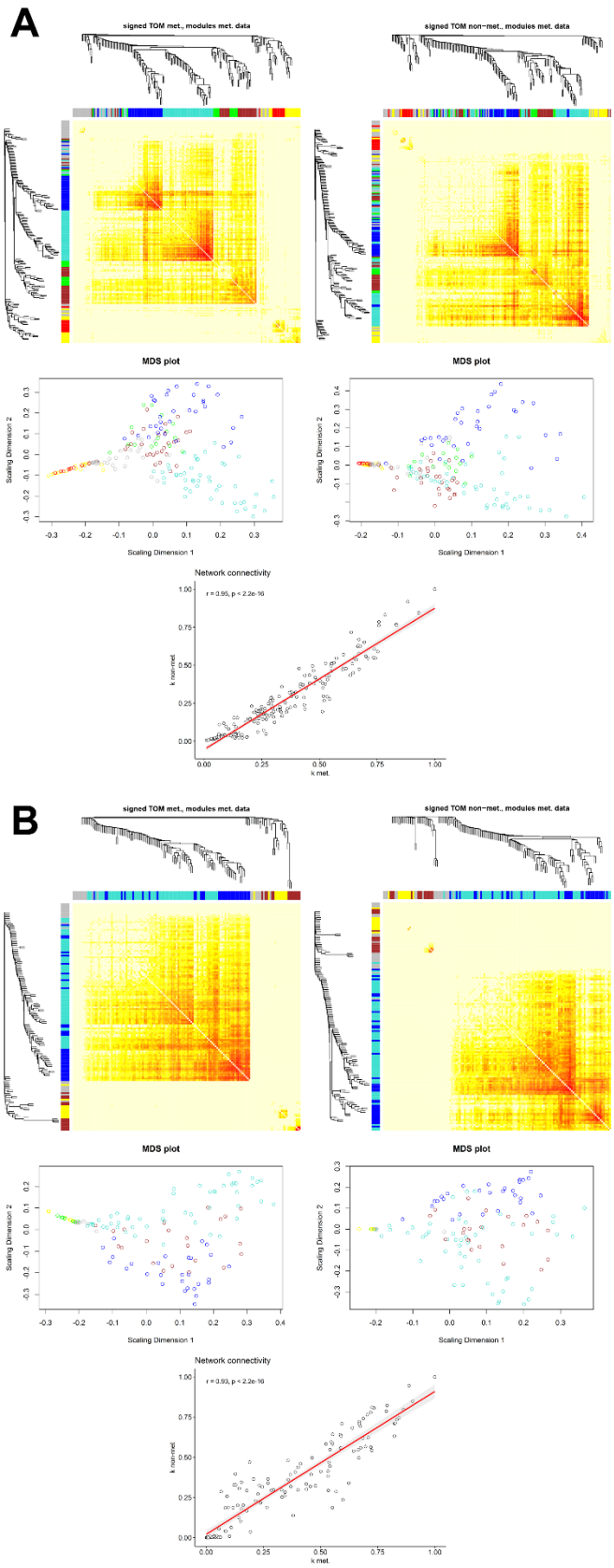

Fig. 2 – Graphical representation of the co-expression networks. Panel A: TOM heatmap with hierarchical clustering results for the metastatic (top left) and non-metastatic (top right) tumor center data; multidimensional scaling (MDS) scatter plot showing the node distribution associated with the

module color attributes for the metastatic (center left) and non-metastatic (center right) tumor center data. The high Pearson's correlation value between the network connectivity k of the metastatic and non-metastatic tumor center data confirms that the two networks are characterized by similar global topological properties. Panel B: Analogous results for the data associated with the tissue sections at a distance of 10cm from the tumor. Figures at left represent the results from the metastatic samples, and figures at right represent the results from non-metastatic samples.

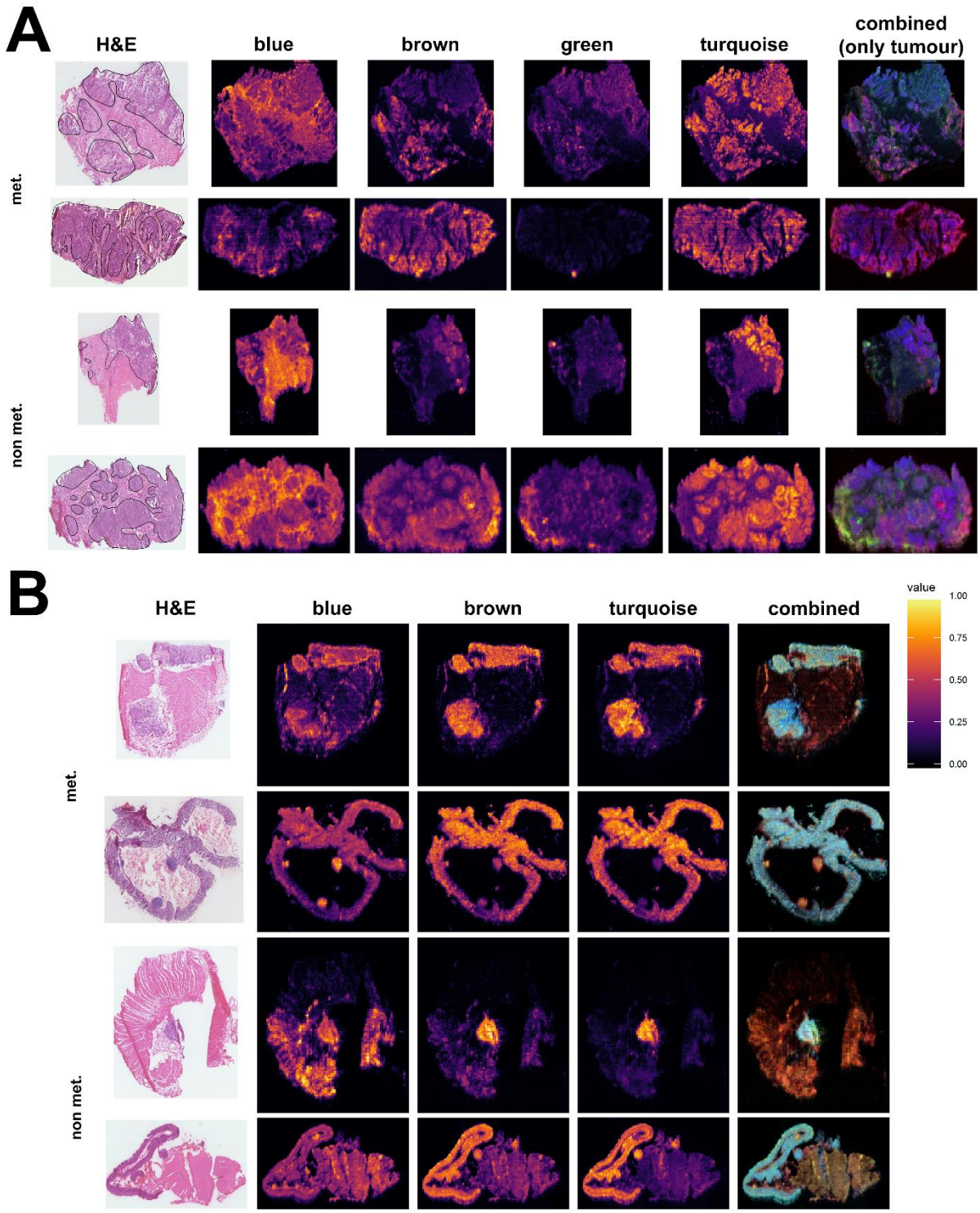

Fig 3 – Spatial visualization of the tissue related module eigenmetabolites (ME). Panel A: results for the tumor core samples. The four tissue related modules are localized in different regions of the cancerous tissue, the "blue" module is mainly expressed in the connective tissue surrounding the tumor (delineated by a black line in the H&E image in the first column), whereas the other three modules mostly involve the tumor. The tumor molecular heterogeneity is captured by the modules "brown", "green" and "turquoise" as shown by the combination of their ME in the last column. Here the three ME expressions are scaled in [0, 1] and visualized as the intensities of RGB channels. Two examples are reported for both the metastatic and non-metastatic samples. Panel B: Analogous results for the tissue

section at 10cm from the tumor center. Here the three tissue related modules are mainly expressed in the epithelial tissue, as shown by the H&E in the first column. The analogous combination of the ME into a RGB image shows the molecular heterogeneity captured by WGCNA.

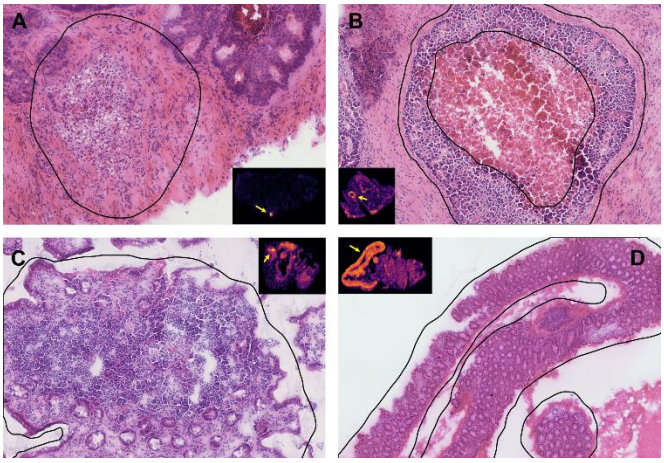

Fig. 4 – Examples of the observed tissue associated with the identified modules. In the metastatic tumor center (A), the weakly preserved "green" module ions are localized in the stroma surrounding the tumor. Noticeably macrophages infiltrating the area are visible together with free fat droplets. The latter have been previously associated with an inflammatory response to the infiltration of the tumor cells (42). In the non-metastatic tumor core, the module ions are still associated with an inflammatory condition that seems to be driven by a necrosis process, as shown in the zoomed H&E stained image corresponding region of the module eigenmetabolites image pointed by the arrow (small box) (B). These observations suggest that different mechanisms are driving the local inflammatory response. Analogously, at a distance of 10cm from the tumor, the weakly preserved "brown" module is localized in sub regions of the metastatic related epithelium involving groups of lymphatic cells as shown by the highly dense cellular populations in the zoomed H&E stained image corresponding to the region pointed by the arrow in the module eigenmetabolites image (small box) (C). A similar result can be observed in the non-metastatic related tissue (D), as represented by the zoomed region of H&E corresponding to the higher intense module eigenmetabolites region. Here, clusters of lymphatic cells are involved in the selected regions of the non-metastatic tissue as well.

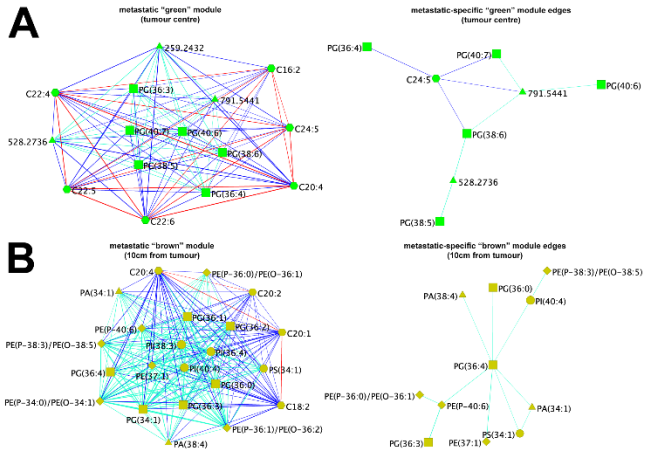

Fig. 5 – Graphical representation of the metastatic related modules that are weakly preserved in the non-metastatic networks. The "green" module associated with the tumor core tissue (panel A, left) consists of PG and PUFA. The edges connecting the PUFA and the PG are colored in blue, whereas those connecting PUFA to PUFA are colored in red. A metastatic edge specificity larger than 0.8 reveals that PG and C24:5 edges are present in the metastatic network but absent in the non-metastatic network (A, right). Moreover, the higher number of edges involving C24:5 and the ion 791.5441 m/z makes them the most representative ions of the metastatic-related module. Analogously, the weakly preserved "brown" module of the network associated with the metastatic 10cm tissue sections reveals a more complex co-localization pattern involving different families of phospholipids and fatty acids (panel B, left). In particular, the metastatic-specific module edges ions, in this case, involve mainly phospholipids, with a central role played by PG(36:4) (panel B, right).