Two-Phase and Graph-Based Clustering Methods for Accurate and Efficient Segmentation of Large Mass Spectrometry Images

Dexter, Alex; Race, Alan M.; Steven, Rory T.; Barnes, Jennifer R.; Hulme, Heather; Goodwin, Richard J. A.; Styles, Iain B.; Bunch, Josephine

DOI:
10.1021/acs.analchem.7b01758

Document Version
Peer reviewed version

Citation for published version (Harvard):
Dexter, A, Race, AM, Steven, RT, Barnes, JR, Hulme, H, Goodwin, RJA, Styles, IB & Bunch, J 2017, 'Two-Phase and Graph-Based Clustering Methods for Accurate and Efficient Segmentation of Large Mass Spectrometry Images', *Analytical Chemistry*, vol. 89, no. 21, pp. 11293-11300.
https://doi.org/10.1021/acs.analchem.7b01758

Two-phase and graph based clustering methods for accurate and efficient segmentation of large mass spectrometry images

Alex Dexter[1,2], Alan M. Race[2], Rory T. Steven[2], Jennifer R. Barnes[3], Heather Hulme[3,4], Richard J.A. Goodwin[3], Iain B. Styles[5], Josephine Bunch[2,6*,]

[1]PSIBS Doctoral Training Centre, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom; [2]National Physical Laboratory, Teddington, Middlesex TW11 0LW, UK; [3]AstraZeneca, Drug Safety and Metabolism, Cambridge CB4 0WG, UK; [4]University of Glasgow, University Avenue, Glasgow, G12 8QQ; [5]School of Computer Science, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom; [6]School of Pharmacy, University of Nottingham, Nottingham, Nottinghamshire NG7 2RD, UK

E-mail: josephine.bunch@npl.co.uk.

Abstract

Clustering is widely used in MSI to segment anatomical features and differentiate tissue types, but existing approaches are both CPU and memory-intensive, limiting their application to small, single datasets. We propose a new approach that uses a graph-based algorithm with a two-phase sampling method that overcomes this limitation. We demonstrate the algorithm on a range of sample types and show that it can segment anatomical features that are not identified using commonly employed algorithms in MSI, and we validate our results on synthetic MSI data. We show that the algorithm is robust to fluctuations in data quality by successfully clustering data with a designed-in variance using data acquired with varying laser fluence. Finally, we show that this method is capable of generating accurate segmentations of large MSI datasets acquired on the newest generation of MSI instruments, and evaluate these results by comparison with histopathology.

Introduction

Mass spectrometry imaging (MSI) is a label free molecular imaging technique capable of spatially mapping molecules in a sample.[1] Typically MSI is performed on biological samples such as thin tissue sections.[2] Since there are a huge number of different molecules in biological tissues, computational analysis of the MSI data is required to mine the large amount of data generated in these experiments[3], and a common task is spatial segmentation which is often performed using methods such as *k*-means or hierarchical clustering.[4,5] These algorithms partition the pixels of the image into groups based on some measure of similarity between them. This enables the different categories of spectra within the data to be identified and separated, thereby segmenting features such as different anatomical structures, or distinguishing between tumour and non-tumour tissues.[6,7]

A number of clustering algorithms have been applied to MSI data. In early studies, *k*-means and agglomerative hierarchical clustering were used to segment anatomies in rat tissue.[4] Recently

algorithms were developed by Alexandrov *et al.* to overcome the pixel-to-pixel variability observed in MSI data and provide a more accurate segmentation of coronal mouse brain images.[8] Since then, the more efficient bisecting *k*-means hierarchical clustering has been used to cluster large MSI datasets including 3D MSI images.[9]

There are a number of underlying difficulties in clustering of MSI data. Since a full mass spectrum is acquired at each pixel location, the data is high-dimensional and the distance metrics that are typically used to measure inter-pixel similarity converge to the same value and are not useful for discrimination.[10] Consequently, clustering on high dimensional data is often unable to converge or falls into local minima, resulting in poor segmentation. In addition to this high dimensionality, MSI data suffers from a high degree of inter-pixel variability, stemming from a number of sources, and tissue regions with very similar molecular composition may yield spectra with significant differences. Despite these challenges, clustering algorithms used in MSI have been shown to be able to distinguish between different anatomies, tumour types, and even distinguish intra tumour heterogeneity.[5,6,11] In order to be clinically applicable, spatial segmentation of pathological regions must be extremely reliable and accurate since poor accuracy would ultimately lead to poor diagnostic accuracy and the possibility of incorrect treatment or intervention. For simple applications such as segmentation of tissue from matrix, or of highly differentiated tissues, simple clustering algorithms such as *k*-means and hierarchical clustering are sufficient.[5,6,12] However when a larger number of fairly similar anatomies need to be segmented, the performance of these algorithms decreases significantly.[8] This can be overcome by the use of more sophisticated algorithms but this comes with increased computational cost, and may require additional dimensionality reduction steps.[8]

Graph theory-based clustering algorithms are frequently used in image segmentation and have been shown to produce more accurate clustering results than simple algorithms such as *k*-means.[13-15] In graph based clustering, the data is represented as a graph in which each pixel is represented as a node in the graph, and the graph edges represent the similarity of the connected

nodes. The graph can be represented by its connectivity or similarity matrix M which can be constructed in one of two ways. In the first, the matrix is constructed such that its elements M_{ij} represent the similarity of the spectra at pixels *i* and *j* by some measure. [16] In the second method, $M_{ij} = 1$ if the spectra at pixels *i* and *j* are within a specified threshold similarity, or $M_{ij} = 0$ otherwise. The clustering itself is then performed using any algorithm (usually *k*-means) on a the number of the eigenvectors specified in table 1 of this connectivity matrix, selected in order of their eigenvalues (from second smallest to largest). The eigenvectors of this similarity matrix allow the optimal partitions of the pixels that maximally preserve connectivity within these data. Three of the most widely used algorithms that use this approach are spectral clustering which clusters the eigenvectors by the *k*-means algorithm[16], minimum cuts which bipartitions the graph such that the connection between the two subsets is minimal[17], and normalised cuts which bipartitions the graph based on some threshold of the eigenvector with the second smallest eigenvalue[14]. Since the *k*-means clustering or bipartition step is performed on only a few eigenvectors (or even a single one), this potentially alleviates some of the issues associated with high dimensional data, and the spectral clustering is related to non-linear dimensionality reduction. The main use of graph based clustering has been in image segmentation[14,18], and offers an alternative method for accurate segmentation of MSI datasets in clinical analysis. However, in graph cuts algorithms, it is normally necessary to specify a number of nearest neighbours in order to construct the connectivity matrix to reduce memory requirements. This added variable can be difficult to select, without *a priori* knowledge of the data as the number of nearest neighbours should be approximately equal to the expected cluster size.[19] Since MSI is often used as an exploratory technique this will not always be known. This can be alleviated by forming a fully weighted graph where the connectivity of two spectra is weighted by their spectral similarity as determined by an appropriate distance metric. This however imposes additional memory requirements as the similarity between every pixel and every other pixel must be calculated and stored.

As well as issues with accuracy of clustering, new developments in instrumentation mean that the size of MSI datasets are continually increasing, both in terms of the number of spectral channels observed and in term of the number of pixels in a given dataset.[20,21] This is a significant problem because datasets are becoming too large to fit into the available RAM of a standard PC, and even high performance computers, limiting the ability to routinely perform multivariate analysis. This usually necessitates the reduction of the data, either to minimal peak lists or using multivariate dimensionality reduction methods such as principal component analysis (PCA), random projection or t-distributed stochastic neighbour embedding (t-SNE).[3,6,22] An alternative approach to handling large data used in many other fields is the divide and conquer approach.[23] This works by recursively breaking a problem down into smaller pieces that can be dealt with easily. These pieces can then be recombined to solve the original problem. This approach is used to efficiently solve numerous problems in computer sciences, from data sorting[24] to the fast Fourier transform[25]. For example, the two-phase *k*-means clustering algorithm uses this principle to group items from large databases that cannot be stored in RAM.[26] The basic algorithm is as follows;

1) Assign the data to one of a predetermined number of subsets *s*
2) Load in one subset and apply necessary preprocessing
3) Cluster the subset into *k* groups using *k*-means
4) Use the cluster centroids to form a compressed representation of this subset
5) Combine all cluster centroids into one dataset and cluster the compression set into *k* groups using *k*-means.
6) Propagate clustering assignments back to original data

For full details of the algorithm see supporting information algorithm S2.

Using this method significantly reduces the RAM required to perform clustering, since only a small subset needs to be in memory at any one time. Since MSI data is both large and high dimensional, we combine graph-based clustering with the two phase divide and conquer approach to accurately cluster large MSI datasets. We demonstrate superior segmentation results using these

algorithms on coronal, sagittal and transverse brain datasets. Following this, we validate these results on synthetic MSI data, and finally go on to demonstrate their application to a large MSI dataset from mouse colon.

Experimental section

Mass spectrometry imaging

Coronal, and transverse sections were obtained from mouse brain, sagittal from rat brain. All sectioning was carried out at 12 µm thickness and thaw mounted onto either glass slides (Superfrost, Thermo Fisher Scientific, Waltham, MA USA) for mouse brain, or stainless steel MALDI target plates for sagittal rat brain. Following mounting, samples were coated with CHCA matrix (5 mg/mL, 80% MeOH 0.1% TFA) using an automated pneumatic sprayer (TM-sprayer, HTX imaging, Chapel Hill, NC, USA). Coronal and transverse MALDI images were acquired using a Synapt G2Si (Waters, Manchester, UK), using a pixel size of 45 x 45 and 30 x 30 µm respectively, over an *m/z range* of 100-1200 Da. Full experimental details on the sagittal rat brain data acquisition was described by Carter *et al*.[27]

Full details of the variable fluence experiments have been described previously.[28] Briefly, serial coronal mouse brain sections were thaw mounted onto a single stainless steel MALDI plate. These were then coated in CHCA as above. Mass spectrometry images were then acquired using a QSTAR XL Qq-ToF mass spectrometer fitted with an oMALDI II ion source (Sciex, Warrington, UK) in continuous raster mode. Data from the second to sixth sections using the 100 µm diameter round core fibre were used as described in the study by Steven *et al.*[28]

Synthetic data were generated using a statistical modelling approach, modelling clusters of MSI data as multivariate normal distributions from the reference data as described by Dexter *et al*.[29]Synthetic datasets with three anatomical regions (from brain stem, lateral septal complex, and isocortex) were generated with between 3,000 and 300,000 pixels in increments of 3,000 pixels, with equal numbers of pixels from each region. The large synthetic brain MSI image was generated based

on the original masks from Dexter *et al.*[29] scaled up by a factor of 3 in *x* and *y* to give a total of 187,425 pixels.

Mouse colon samples were collected, prepared using the 'Swiss Roll' technique[30] and embedded in 2.5% carboxymethyl cellulose (Sigma-Aldrich) in sterile water. Full details on sample preparation can be found in the supporting information. High spatial resolution MS images were acquired using a RapiFlex MALDI ToF/ToF (Bruker Daltonics, Germany) in reflectron positive ion mode, using a pixel size of 5 x 5 µm, over an *m/z* range of 200-1000 Da.

Data processing and analysis

Data processing was performed on an Intel Xeon quad core CPU E5-2637 v2 (3.50 GHz) with 64 GB of RAM. All data were converted from proprietary format to the mzML format using the *msconvert* tool in the ProteoWizard[31] software, and then into imzML using imzMLConverter[32]. This was then imported into MATLAB (version R2014a and statistics toolbox, The Math-Works, Inc., Natick, MA, USA) using the SpectralAnalysis software package[33]. QSTAR data were zero-filled using the QSTARZeroFilling routine in SpectralAnalysis, followed by three iterations of Savitzky-Golay smoothing with a window size of 7 and second order polynomial, and the negative signals produced by the smoothing were then removed by truncating to zero. The data acquired on the Synapt were zero filled using the interpolated rebinning function in SpectralAnalysis with a bin size of 0.01 and no smoothing applied. Total spectra were then generated from each dataset, peak picked using the gradient method, and the peak intensities were extracted for individual pixels. *k*-means clustering was performed using the function *kmeans* from the Matlab Statistics toolbox using the parameters given in the Results and Discussion section with three replicates and random starting clusters. The spectral clustering algorithm (SI Algorithm S1) was used in all cases of graph based clustering, and the full weighted similarity graph representation was used in each case where the edges were represented by cosine similarity.

Data partitioning for the different subsets in two-phase *k*-means and two-phase graph cuts clustering was performed by pseudo-random assignment of each pixel into a predefined number of

subsets such that the subsets were of equal size, and all spectra were assigned to a subset. The subset sizes were 17,000 spectra for the synthetic data, 25,000 spectra for the transverse brain data, and 19,000 for the gut data. In all cases, the cosine similarity measure was used for weighted graph construction and *k*-means clustering based on previous literature on distance metric choice in MALDI MSI.[29]

When clustering the smaller datasets using graph cuts, *k*-means clustering was performed on the smallest 250 eigenvectors of the connectivity graph. For the two-phase graph cuts of the large synthetic brain data, *k*-means was performed on the smallest 500 eigenvectors of the connectivity graph of the subsets, and the smallest 20 eigenvectors of the graph of the compression set. For the two-phase graph cuts of the transverse brain data, *k*-means was performed on the smallest 600 eigenvectors of the connectivity graph of the subsets, and the smallest 10 eigenvectors of the graph of the compression set. In the two-phase graph cuts of the mouse colon data, *k*-means was performed on the smallest 600 eigenvectors of the connectivity graph of the subsets, and the smallest 10 eigenvectors of the graph of the compression set. For a summary of these parameters see table 1. The number of eigenvectors selected represent approximately 2% of the total number of eigenvectors present. A range of values were investigated and this was found to produce the best segmentation based on visual inspection. A method for a more objective selection of this value would make an interesting topic for future research. The subset sizes were chosen based on a compromise between efficiency and having subsets that were representative of the whole dataset.

Full details on the spectral clustering, two phase *k*-means and two phase graph cuts algorithm see supporting information.

Table 1 Two-phase graph cuts parameters used

Dataset	Eigenvectors used to cluster subsets	Eigenvectors used to cluster compression set	Subset size (pixels)
Large synthetic brain	500	20	17,000
Transverse brain	600	10	25,000
Mouse colon	600	20	19,000

Results and Discussion

To evaluate performance of different clustering approaches, MALDI MSI data from coronal and sagittal murine brain sections were processed using k-means, agglomerative clustering, graph cuts and bisecting *k*-means algorithms. A summary of the results from this comparison is shown in Figure 1 alongside images from the Allan brain atlas from the same anatomical location in the brain. Using graph cuts clustering on MSI images of coronal and sagittal brain produces much clearer anatomical segmentation based on a visual comparison with the Allen brain atlases[34] when compared to other clustering methods such as *k*-means and hierarchical clustering (Figure 1). In the sagittal brain data, only the graph cuts algorithm is able to clearly segment caudate putamen (turquoise), cerebral cortex (orange and grey), thalamus (light green), midbrain (cream and blue) and hippocampus (purple). Similarly, in the coronal brain data, only the graph cuts segmentation separates the isocortex (green) from olfactory areas (purple), and identifies the caudoputamen (dark blue) and brain stem (red) areas. Additional results from *k*-means and graph cuts clustering with different values of k are provided in the supporting information figure S1.

This gives a good initial indication that graph cuts based clustering can accurately segment MSI data. These initial results are not readily generalisable as it cannot be ruled out that the inherent characteristics of these datasets are more favourable to this approach and more controlled experiments are required. We perform several experiments to test the applicability of the proposed methods across different datasets.

In a clinical setting, the result of the clustering must be robust to any noise or spectral differences in the data that result from the pixel to pixel variability derived from experimental sources within an MS image such as the effects of inhomogeneous sample preparation and laser instability.[35,36] In order to analyse a dataset with a controlled and known reduction in the spectral quality of the data, a series of mouse brain datasets acquired at decreasing laser fluence were studied. The spectral quality of these data decreases as the fluence falls below the threshold for ionisation (Figures S2– S6).[28] These data present an extreme, but controlled, example of variable

quality spectra – in this case due to decreasing laser energy. This reflects the variability within MSI data, where artefacts, e.g. those from inhomogeneous matrix deposition, cause localised deviations in spectra quality. In this situation, the graph cuts clustering algorithm is visibly superior to *k*-means clustering at segmenting the anatomical features in the tissue (Figure 2). This makes it more suitable for use when anatomical segmentation is the desired result of the clustering, for example in biomedical imaging applications. This result can be attributed to the preservation of connectivity when using the graph cuts clustering algorithm. The data acquired at 35 Jm^{-2} will be similar to that acquired at 51 Jm^{-2}, which will in turn be similar to that acquired at 78 Jm^{-2} and so on. Therefore there is a continuous path of connectivity between the data acquired at the lowest and highest fluences. If the connectivity is broken, as would happen if only the lowest and highest fluence datasets are clustered, the graph cuts algorithm is able to distinguish between these experimental variances (Figure S7). Therefore, in studies where there is likely to be an incremental changes within the data, graph cuts clustering should be used when the desired result is to ignore these incremental changes. If the desire is to segment and identify these incremental changes, then the *k*-means clustering algorithm is more suitable.

In studies where *k*-means is to be used, the memory requirement and speed of this algorithm can be improved through the use of the two-phase approach. The time complexity of *k*-means clustering is $O(n.d.k.i)$ where *n* is the number of pixels, *d* the number dimensions, *k* the number of clusters and *i* the number of iterations.[37] The number of iterations, however, can itself increase exponentially with *d* and *k*.[38] Using the two-phase *k*-means approach, clustering is only performed on a small subset of the whole data thereby reducing the complexity by reducing both *n* (directly) and *i* (indirectly). The two-phase *k*-means clustering is at its most memory efficient when the subset and compression set are of equal size and at this point, time complexity scales as $O\left(s.\sqrt{n}\,.d.k.i\right)$ where *s* is the number of subsets used. While this may not initially appear to be an improvement, since the number of iterations *i* scales exponentially with number of pixels *n,* the improvements seen in time complexity are increasingly significant as *n* increases. To demonstrate

these improvements, synthetic MSI datasets consisting of 3 regions (from brain stem, lateral septal complex, and isocortex), with a varying number of pixels (3000 to 300,000) and 8,193 mass channels were generated using a statistical modelling approach[29] and clustered using *k*-means and two-phase *k*-means clustering. The time taken to perform *k*-means clustering was around three times greater than two-phase *k*-means (Figure S8) in all cases, with no significant difference in accuracy (as measured by Rand indices close to 0.9) when using two-phase *k*-means clustering (Figure S9). In addition to the reduction in time complexity, the memory requirement for two-phase *k*-means scales by $\sqrt{n \times k}$. However, since $k<<n$, this method scales much more efficiently than the standard *k*-means algorithms. The memory requirements of both *k*-means and bisecting *k*-means algorithms scale linearly with number of pixels (requiring the full data to be stored in memory). This becomes increasingly important as *n* increases, such as in high spatial resolution or 3D MSI images, where the data becomes too large to store in RAM.[9,21,39] The main issue with this is that as the number of pixels increases with newer developments in instrument design, the number of mass channels that can be retained decreases, often requiring reduction of spectra to minimal peak lists that may lose critical low intensity features in the data (Figure S10).[6,22] For example, given a PC with 8 GB of RAM, and a dataset with 1,000,000 pixels, 1,000 peaks can be retained when loading these data into memory before considering any processing.

The alternative to this is to perform dimensionality reduction via methods such as PCA or random projection, however this then requires additional computation, and in other fields has been shown to degrade cluster quality in some cases.[40] Using the two phase clustering methods allow even the largest of MSI datasets to be clustered without having to compromise on the information retained in the data (Figure S11). It is worth noting however that the two-phase algorithms assume that the subsets used for clustering are representative of the full dataset. Therefore, larger subsets will generally produce more accurate clustering results and therefore subset size should be chosen based on the available RAM to the user.

While the graph cuts clustering algorithm gives clearer anatomical segmentation, it is important to quantify these improvements in order to give a non-subjective measure of how this clustering performs. In any biological sample there will always be inherent unknowns, preventing any quantitative analysis of these results. Recently, statistical modelling has been shown to be capable of producing datasets with known spatial distributions, suitable for quantitative evaluation of clustering in MSI.[29] Therefore, in order to evaluate the different clustering algorithms with respect to one another, a synthetic dataset comprising of 7 regions and 20,825 pixels was generated using the statistical modelling method. This was clustered using some of the existing algorithms in the MSI literature (*k*-means, bisecting *k*-means, and agglomerative hierarchical) as well as the new graph cut-based method. In all cases 7 clusters were used, with the cosine similarity, and in the case of the *k*-means based methods, 3 replicates were used. The results of the clustering were then evaluated using the Rand index.[41] Graph cuts clustering was found to outperform all other clustering algorithms with indices of > 0.9 compared to < 0.7 for the other algorithms (Figure S12).

While the graph cuts clustering produces better clustering results than existing clustering algorithms, this comes with an increased computational cost reducing its effectiveness as an algorithm for clustering large MSI data. In order to perform graph cuts clustering, the full pairwise distance matrix must be calculated, along with the eigenvectors of this distance matrix. In a dataset with *n* pixels and *d* dimensions, the time complexity of this scales as $O(n^2.d)$ and the memory required for this with n^2. This becomes increasingly intractable as *n* increases, rapidly approaching memory requirements beyond even the most powerful processing PC (Figure 3).

As with two-phase *k*-means clustering, by using a two-phase clustering approach, the memory requirement and time complexity required to perform graph cuts clustering can be significantly reduced. In two-phase graph cuts clustering, the pairwise distances matrices are calculated on the smaller subsets, thereby reducing the complexity to O(*n.d*) and the memory requirement now scales linearly with the number of pixels. This means that even the largest of MSI

datasets can be clustered using this method on a standard desktop PC (Figure 3). For a full analysis of the complexity of the proposed method can be found in the supporting information.

To test the accuracy of the two-phase graph cuts algorithm on large datasets, a large synthetic dataset comprising 7 regions totalling 187,425 pixels and 8,193 mass channels was generated using statistical modelling. This would require 11.4 GB to load into RAM, within the capabilities of some higher end PC's but not all standard PC's or laptops. Additionally, in order to calculate and store a full pairwise distance matrix for this dataset would require over 260 GB of RAM, well beyond even high performance PC's. While large, this dataset is still well below the size of datasets often acquired on newer generation instruments or large 3D MSI datasets.[9,21,42] This dataset was clustered using *k*-means, two-phase *k*-means and two-phase graph cuts clustering algorithms. The two phase *k*-means clustering produced almost identical results to the standard *k*-means algorithm (Figure 4), but the clustering required 1.5 GB RAM for two-phase *k*-means vs 11.5 GB for standard *k*-means. Additionally, the two-phase graph cuts clustering produces much more accurate results than both the *k*-means and two-phase *k*-means, as measured by the Rand index (Figure 4), while still requiring less RAM than the *k*-means clustering algorithm (< 3 GB).

Two-phase graph cuts clustering was then applied to a large MS image of a transverse mouse brain acquired with a pixel size of 30 µm comprising of 101,390 pixels. This represents both a large number of pixels (> 100,000), rich and complex lipid spectra (>7,000 peaks), and a large number of image features (> 10 anatomical regions). As with the smaller image from the coronal mouse brain image, the two-phase graph cuts clustering produces a clearer anatomical segmentation than two-phase *k*-means clustering with respect to the expected anatomies (Figure 5).

A larger dataset from an MSI image of gut tissue acquired with a pixel size of 5 µm was also segmented by two-phase graph cuts, and by two-phase *k*-means clustering. This dataset contained 400,625 pixels, and 6,886 spectral channels, is too large to load into RAM on a standard PC (>20GB), and would require >1TB memory to store a full pairwise distance matrix. In addition to MSI analysis, histopathological assessment was performed using a haematoxylin and eosin (H&E) stained serial

tissue section (Figure 6A). Four distinct anatomical layers are readily apparent; the mucosa, the sub-mucosa, the muscularis propria (externa) and the serosa. The mucosa (red) represents the innermost layer of the colon and can be sub-divided further into the epithelium, a supportive lamina propria and an outer muscularis mucosae. The mucosal epithelial layer is formed from tightly packed glands (or crypts) that open onto the surface epithelium. The neck of the glands are lined by absorptive epithelial cells, goblet cells and enteroendocrine cells, whereas stem cells and transit amplifying cells are located towards the base of the glands. The sub-mucosa (green) lies directly beneath the mucosa. The muscularis propria (grey) surrounds the sub-mucosa and consists of the inner circular and outer longitudinal smooth muscle layers. The outermost layer, the serosa (blue), consists of a thin layer of connective tissue lined by a single layer of mesothelial cells forming the visceral peritoneum. It is important to note that due to the orientation of the colon within the Swiss roll, minor region differences in the plane of the tissue are evident within the section. In addition, in some areas, the serosa and muscularis propria in particular are variably intact.

Comparison of the area of the H&E section analysed and the two-phase k-means clustering clearly demonstrates that the clustering provides little or no discrimination between the various anatomical regions of the colon (Figure 6B). In contrast, the two-phase graph cuts method can discriminate tissue from ‘non-tissue’ and appears to start to identify specific regions (Figure 6C). The differentiation between mucosa (red in H&E stain) and underlying sub-mucosa (green in H&E) / muscularis externa (grey in H&E stain) is particularly clear. Although the limits of resolution do not allow individual cell identification, the appearance of the clustering within the mucosa is consistent with the histological appearance of the mucosa and may partially capture the glandular structure of the epithelium. The slight differences observed in the two-phase clustering between the sections of colon within the Swiss roll may be a consequence of the variability in the section of plane as previous described.

Conclusions

Graph based clustering is shown to produce better anatomical segmentations of MSI data than other algorithms used in these challenging application areas on both synthetic and experimental datasets. This segmentation is more robust towards spectral changes caused by experimental factors, provided that the variability maintains spectral connectivity in the data. In cases where the full pairwise distance matrix cannot be stored in memory, or the data itself is too large to load into RAM, the two-phase clustering approach can be used to reduce this cost and speed up the clustering process. This comes with only a very minimal reduction in segmentation performance. With new developments in instrumentation, along with a growing need and capability to combine multiple datasets together, MSI datasets are rapidly growing in size. The algorithms presented in this work provide a means to accurately and efficiently segment the next generation of MSI data. Future research should consider the effect of different subset sizes on the accuracy of the two-phase clustering; how the numbers of eigenvectors affects the clustering results; and the efficiency of these algorithms for segmentation of image data collected using high mass resolution MS instruments where the number of peaks vastly exceeds the numbers handled here.

Acknowledgements

AD gratefully acknowledges financial support from the EPSRC through a studentship from the PSIBS Doctoral Training Centre (EP/F50053X/1), NPL and AstraZeneca. JB, AD, RTS and AMR all gratefully acknowledge funding from NPL Strategic Capability programme 'AIMSHIGHER'. Data supporting this research is openly available from the University of Birmingham data archive at http://findit.bham.ac.uk/.

Supporting information

Details on further experimental data acquisition, derivations for algorithm efficiencies, details on the algorithms used, example spectra from different fluence tissue data, graphs of computational requirements and sensitivities for the different algorithms.

References

(1) Caprioli, R. M.; Farmer, T. B.; Gile, J. *Anal. Chem.* **1997**, *69*, 4751-4760.
(2) Chaurand, P. *J. Proteomics* **2012**, *75*, 4883-4892.
(3) Fonville, J. M.; Carter, C. L.; Pizarro, L.; Steven, R. T.; Palmer, A. D.; Griffiths, R. L.; Lalor, P. F.; Lindon, J. C.; Nicholson, J. K.; Holmes, E. *Anal. Chem.* **2013**, *85*, 1415-1423.
(4) McCombie, G.; Staab, D.; Stoeckli, M.; Knochenmuss, R. *Anal. Chem.* **2005**, *77*, 6118-6124.
(5) Deininger, S. r.-O.; Ebert, M. P.; Fütterer, A.; Gerhard, M.; Röcken, C. *J.Proteome Res.* **2008**, *7*, 5230-5236.
(6) Race, A. M.; Steven, R. T.; Palmer, A. D.; Styles, I. B.; Bunch, J. *Anal. chem.* **2013**, *85*, 3071-3078.
(7) Jones, E. A.; van Remoortere, A.; van Zeijl, R. J.; Hogendoorn, P. C.; Bovée, J. V.; Deelder, A. M.; McDonnell, L. A. *PLoS One* **2011**, *6*, e24913.
(8) Alexandrov, T.; Kobarg, J. H. *Bioinformatics* **2011**, *27*, i230-i238.
(9) Thiele, H.; Heldmann, S.; Trede, D.; Strehlow, J.; Wirtz, S.; Dreher, W.; Berger, J.; Oetjen, J.; Kobarg, J. H.; Fischer, B.; Maass, P. *Biochimica Et Biophysica Acta-Proteins and Proteomics* **2014**, *1844*, 117-137.
(10) Keogh, E.; Mueen, A. In *Encyclopedia of Machine Learning*; Springer, 2011, pp 257-258.
(11) Willems, S. M.; van Remoortere, A.; van Zeijl, R.; Deelder, A. M.; McDonnell, L. A.; Hogendoorn, P. C. *J. Pathology* **2010**, *222*, 400-409.
(12) Abdelmoula, W. M.; Carreira, R. J.; Shyti, R.; Balluff, B.; van Zeijl, R. J.; Tolner, E. A.; Lelieveldt, B. F.; van den Maagdenberg, A. M.; McDonnell, L. A.; Dijkstra, J. *Anal. Chem.* **2014**, *86*, 3947-3954.
(13) Choong, M. Y.; Kow, W. Y.; Chin, Y. K.; Angeline, L.; Teo, K. T. K.; Ieee. *Image Segmentation via Normalised Cuts and Clustering Algorithm*; Ieee: New York, 2012, p 430-435.
(14) Shi, J.; Malik, J. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **2000**, *22*, 888-905.
(15) Jain, A. K. *Pattern Recogn. Lett.* **2010**, *31*, 651-666.
(16) Ng, A. Y.; Jordan, M. I.; Weiss, Y. In *NIPS*, 2001, pp 849-856.
(17) Picard, J.-C.; Ratliff, H. D. *Networks* **1975**, *5*, 357-370.
(18) Zhang, X.; Jiao, L.; Liu, F.; Bo, L.; Gong, M. *Geoscience and Remote Sensing, IEEE Transactions on* **2008**, *46*, 2126-2136.
(19) Von Luxburg, U. *Statistics and computing* **2007**, *17*, 395-416.
(20) Römpp, A.; Guenther, S.; Takats, Z.; Spengler, B. *Anal. Bioanal. Chem.* **2011**, *401*, 65-73.
(21) Ogrinc Potočnik, N.; Porta, T.; Becker, M.; Heeren, R.; Ellis, S. R. *Rapid Commun. Mass Spectrom.***2015**, *29*, 2195-2203.
(22) Palmer, A. D.; Bunch, J.; Styles, I. B. *Anal. Chem.* **2013**, *85*, 5078-5086.
(23) Akl, S. G. *Parallel computation: models and methods*; Prentice-Hall, Inc., 1997.
(24) Cole, R. *SIAM J. Computing* **1988**, *17*, 770-785.
(25) Frigo, M.; Johnson, S. G., The fastest fourier transform in the west; DTIC Document1997.
(26) Pham, D.; Dimov, S.; Nguyen, C. *Proceedings of the Institution of Mechanical Engineers, Part C: J. Mech. Eng. Science* **2004**, *218*, 1269-1273.
(27) Carter, C. L.; McLeod, C. W.; Bunch, J. *J. Am. Soc. Mass Spectrom* **2011**, *22*, 1991-1998.
(28) Steven, R. T.; Race, A. M.; Bunch, J. *J. Am. Soc. Mass Spectrom* **2016**, 1-10.
(29) Dexter, A.; Race, A.; Styles, I.; Bunch, J. *Anal. Chem.* **2016**.
(30) Park, C. M.; Reid, P. E.; Walker, D. C.; MacPherson, B. R. *J. microscopy* **1987**, *145*, 115-120.
(31) Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P. *Bioinformatics* **2008**, *24*, 2534-2536.
(32) Race, A. M.; Styles, I. B.; Bunch, J. *J. Proteomics* **2012**, *75*, 5111-5112.
(33) Race, A. M.; Palmer, A. D.; Dexter, A.; Steven, R. T.; Styles, I. B.; Bunch, J. *Anal. Chem.* **2016**, *88*, 9451-9458.
(34) Lein, E. S.; Hawrylycz, M. J.; Ao, N.; Ayres, M.; Bensinger, A.; Bernard, A.; Boe, A. F.; Boguski, M. S.; Brockway, K. S.; Byrnes, E. J. *Nature* **2007**, *445*, 168-176.

(35) Goodwin, R. J. *J. proteomics* **2012**, *75*, 4893-4911.
(36) Steven, R. T.; Dexter, A.; Bunch, J. *Methods* **2016**.
(37) Nazeer, K. A.; Sebastian, M. In *Proceedings of the World Congress on Engineering*, 2009, pp 1-3.
(38) Arthur, D.; Manthey, B.; Roglin, H. In *Foundations of Computer Science, 2009. FOCS'09. 50th Annual IEEE Symposium on*; IEEE, 2009, pp 405-414.
(39) Klinkert, I.; Chughtai, K.; Ellis, S. R.; Heeren, R. M. *Int. J. Mass Spectrom.* **2014**, *362*, 40-47.
(40) Yeung, K. Y.; Ruzzo, W. L. *Bioinformatics* **2001**, *17*, 763-774.
(41) Rand, W. M. *J. Am. Stat. Assoc.* **1971**, *66*, 846-850.
(42) Römpp, A.; Spengler, B. *Histochem. Cell Biol.* **2013**, *139*, 759-783.

Figures

Figure 1. Comparison of existing clustering algorithms used in MSI, and graph cuts clustering applied to an MSI images of a coronal (*k* = 7) and sagittal (*k* = 20) brain sections as compared to the Allen brain atlas (bottom). Coronal mouse brain data was acquired with 45 x45 µm pixels and contained a total of 20,000 pixels, sagittal rat brain was acquired with 100 x 100 µm pixels and contained 12,500 pixels.

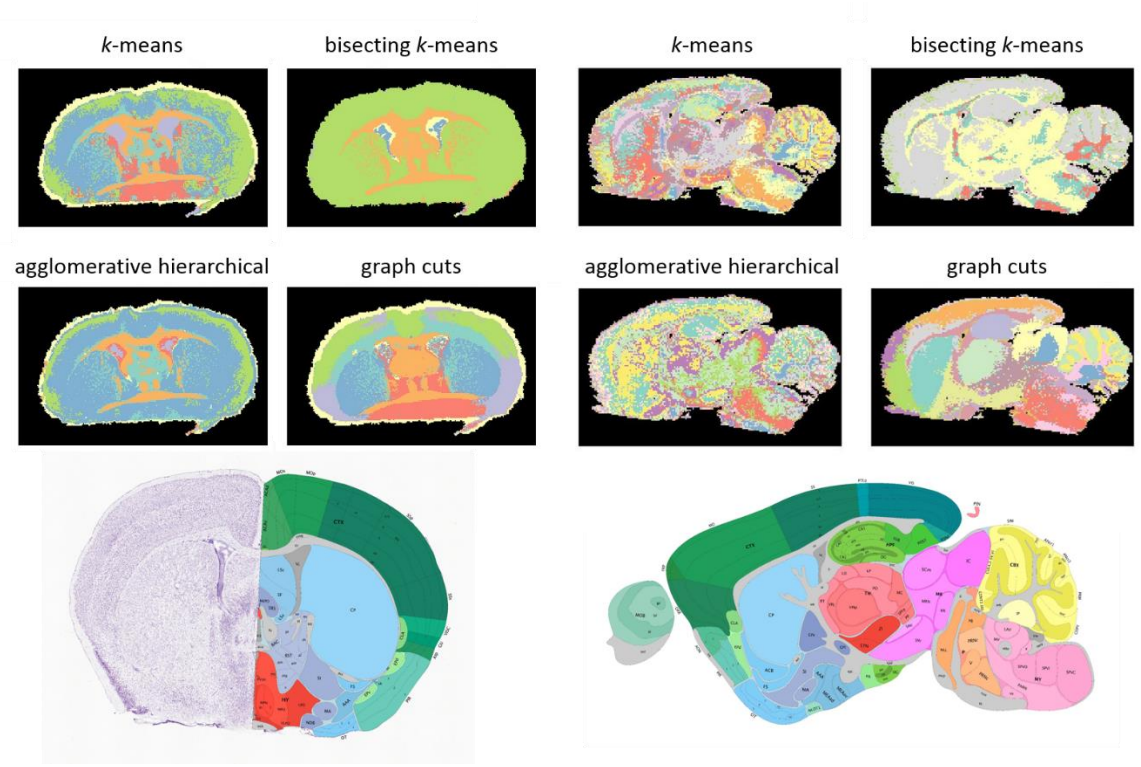


Figure 2. Comparison of graph cuts and *k*-means clustering on data acquired at decreasing laser fluence showing consistent anatomical segmentation with graph cuts, compared to separation of experimental parameters with *k*-means clustering

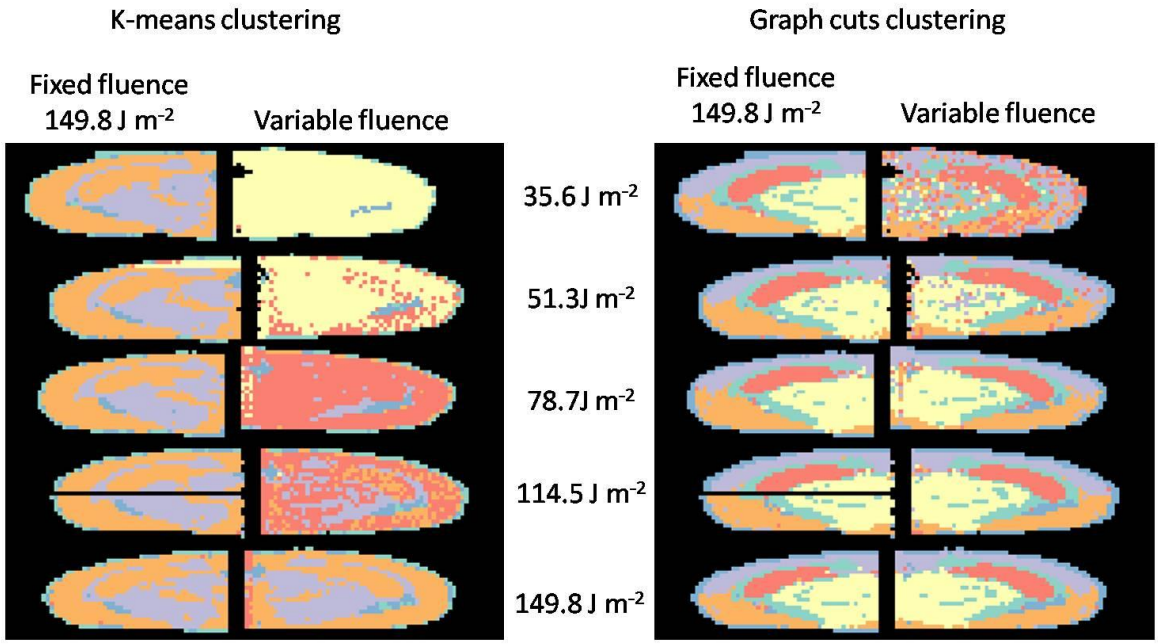


Figure 3. Graph of memory requirements against number of pixels when using the graph cuts and two-phase graph cuts algorithms

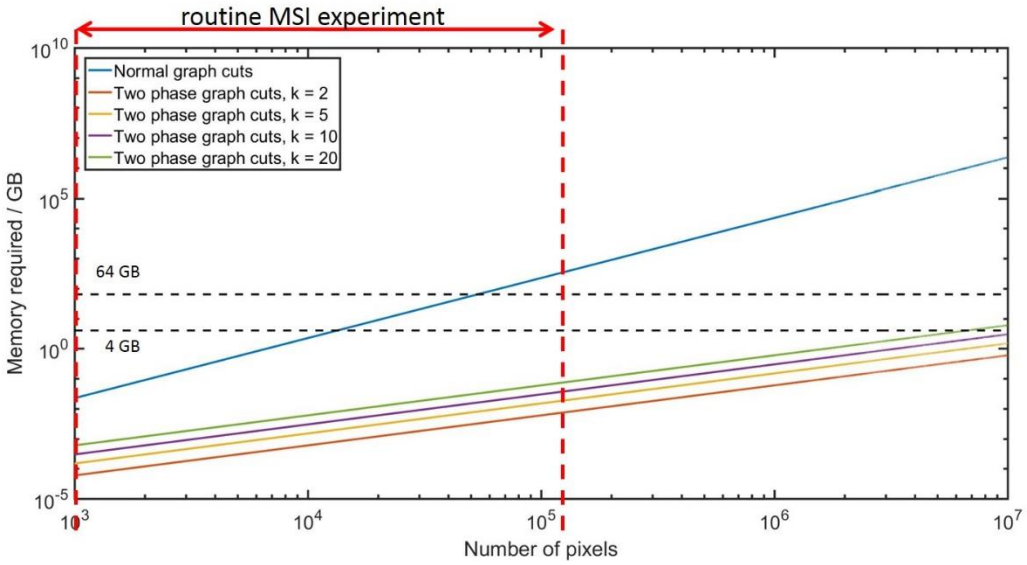


Figure 4. Comparison of two-phase graph cuts, bisecting *k*-means, and k-means clustering on a large synthetic dataset containing seven regions and 187,425 pixels.

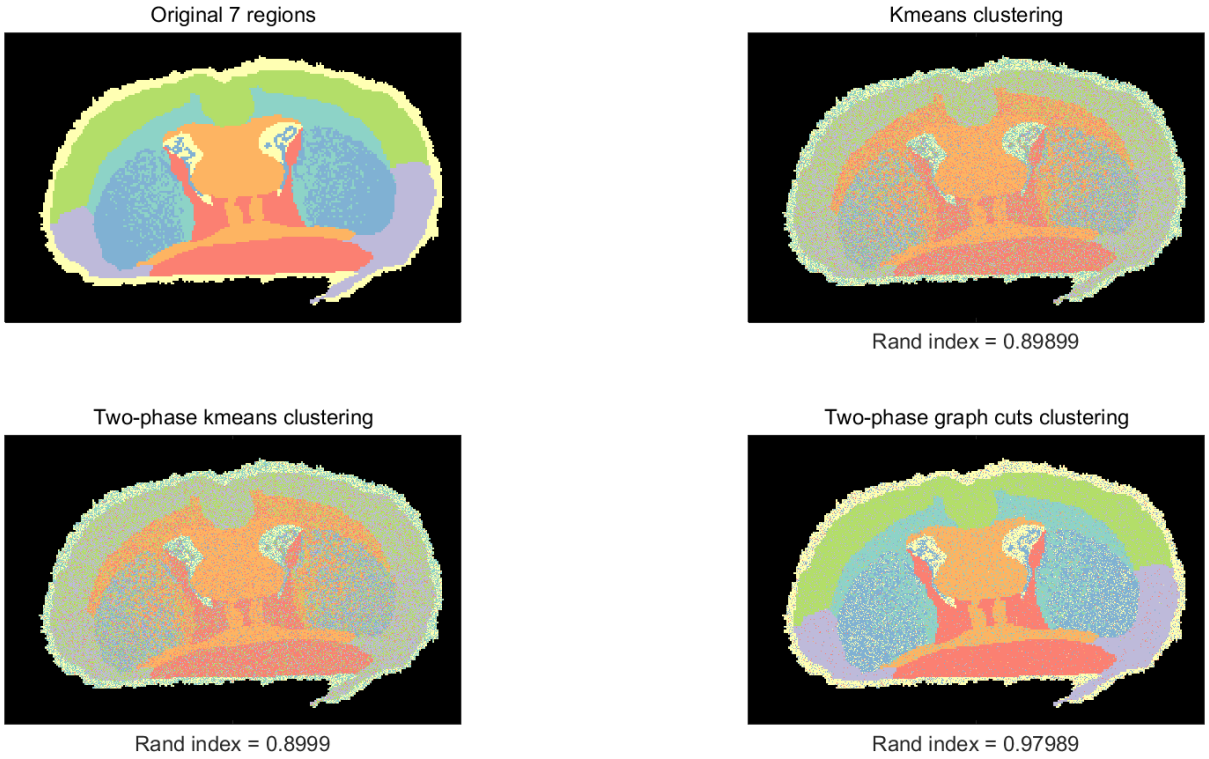


Figure 5. Comparison of two-phase *k*-means and two-phase graph cuts clustering on transverse brain data acquired using 30 x 30 µm pixels, containing over 100,000 pixels.

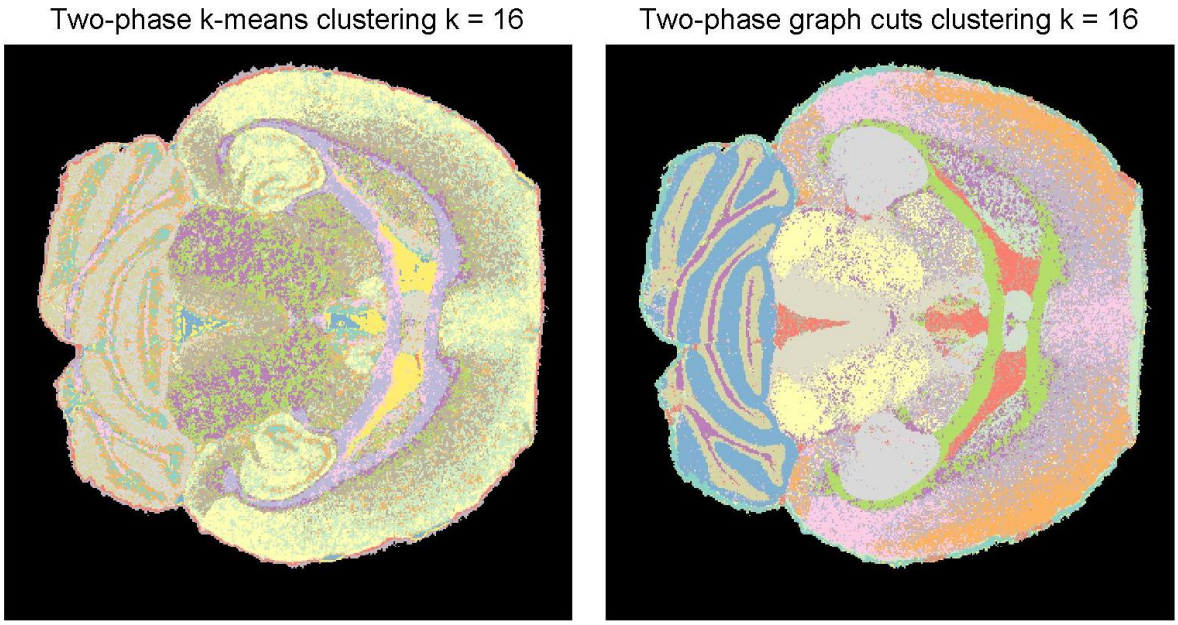
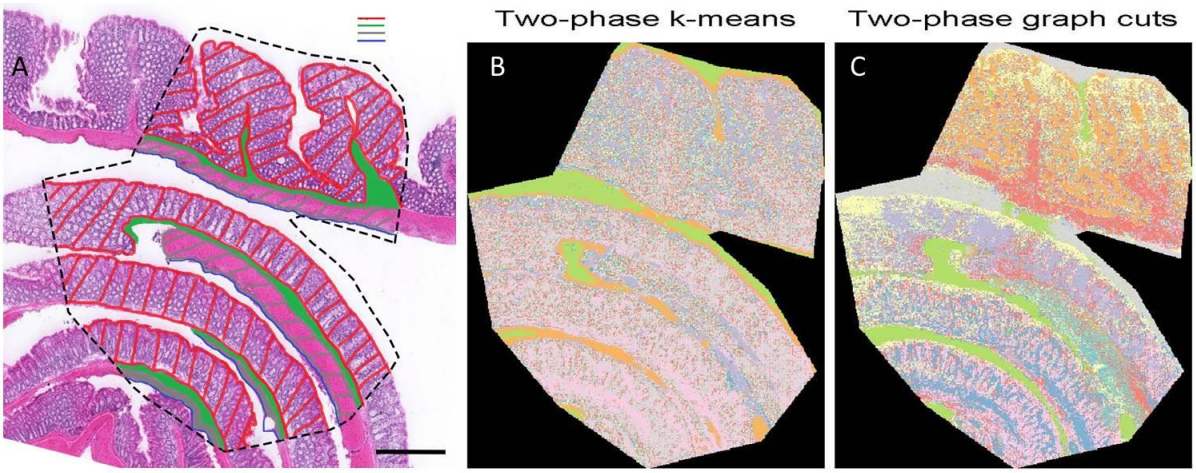
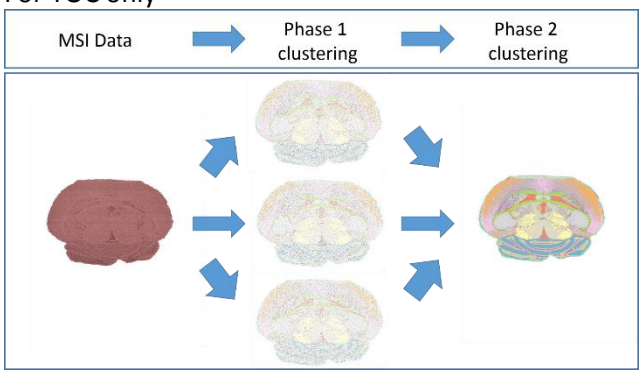


Figure 6. Comparison of two-phase *k*-means (b) and two-phase graph cuts (c) clustering on large gut image data acquired using 5 x 5 µm pixels, containing over 400,000 pixels. Compared to H&E stained image labelled by a pathologist (a) (mucosa in red, sub-mucosa green, muscularis propria grey, and serosa blue). Scalebar in (a) represents 500 µm.



For TOC only



Supporting Information for:

Two-phase and graph based clustering methods for accurate and efficient segmentation of large mass spectrometry images

Alex Dexter[1,2], Alan M. Race[2], Rory T. Steven[2], Jennifer R. Barnes[3], Heather Hulme[3,4], Richard J. A. Goodwin[3], Iain B. Styles[5], Josephine Bunch[2,6*,]

[1]PSIBS Doctoral Training Centre, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom; [2]National Physical Laboratory, Teddington, Middlesex TW11 0LW, UK; [3]AstraZeneca, Drug Safety and Metabolism, Cambridge CB4 0WG, UK; [4]University of Glasgow, University Avenue, Glasgow, G12 8QQ; [5]School of Computer Science, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom; [6]School of Pharmacy, University of Nottingham, Nottingham, Nottinghamshire NG7 2RD, UK

E-mail: josephine.bunch@npl.co.uk.

Contents

Experimental for mouse colon data

Mouse colons were collected, prepared using the 'Swiss Roll' technique and embedded in 2.5% carboxymethyl cellulose (Sigma-Aldrich) in sterile water. The embedded colons were frozen in a slurry of ethanol and crushed dry ice, and then stored at -80°C. The colons were cut to 10 µm sections in a cryostat microtome at -18°C. The sections were cut in a specific order to take 3 sections for histology and 2 sections for MSI. Sections for MSI were thaw mounted on conductive indium tin oxide coated (ITO) slides (Bruker Daltonics, Germany) and sections for histology onto normal microscope slides. Slides were stored at -80°C until imaging or staining. Tissue sections thaw mounted onto ITO slide for imaging were dried in a stream of nitrogen when removed from -80°C storage. Optical images were taken using a standard flatbed scanner (Seiko Epson, Negano, Japan) prior to sample preparation and MALDI matrix application. Sections were treated with 2,4-diphenyl-pyranlium tetrafluoroborate (DPP-TFB) to derivitize endogenous primary amines as previously described.[1] Briefly, DPP-TFB, 9.6 mg was dissolved in 1.2 ml of 100% methanol and sonicated for 10 min and 3.5 µl of trimethylamine was added to 6 ml of 70% methanol. The DPP-TFB solution was gradually added to the 70% methanol and this final solution was sprayed onto slides using an automatic matrix sprayer (TM-Sprayer, HTX Technologies) at 0.08 mL/min, 80°C with 30 passes. The slide was incubated in a chamber with 50% methanol for 15 min, and dried every 5 min under a nitrogen stream.

Two phase *k*-means complexity

Given a dataset of *n* pixels being divided into *k* clusters by the two-phase *k*-means clustering algorithm with *s* subsets. Assuming the mass channels remain constant, the subset of size *n/s* and compression of size *sk* need to be in RAM at any given time. Therefore the total RAM requirement is *n/s* + *sk*. To find the minima of this (where the requirement is lowest) we can differentiate this with respect to the number of subsets *s* to give

(1)

$$\frac{d\left(\frac{n}{s}+sk\right)}{ds}=0$$

Which re-arranges to

(2)

$$-\frac{n}{s^2}+k=0$$

Then

(3)

$$sk=\frac{n}{s}$$

As previously established, *sk* is the compression set, and *n/s* the subset size. The RAM required at this optimal number of subsets will be related to the number of subsets which in this case will be

(4)

$$RAM_{min}=\frac{n}{s}+sk$$

Since at the optimal number of subsets

(5)

$$s=\sqrt{\frac{n}{k}}$$

Then

(6)

$$RAM_{min}=\frac{n}{\sqrt{\frac{n}{k}}}+\sqrt{\frac{n}{k}}k$$

Which re-arranges to give

(7)

$$RAM_{min} = 2\sqrt{nk}$$

Therefore the space complexity of the algorithm scales by $O\left(\sqrt{n.k}\right)$

Two phase graph cuts complexity

For the two-phase graph cuts clustering algorithm, or any algorithm that requires a full pairwise distance matrix calculation, the number of mass channels *d* will also affect the optimum number of subsets, and there are two possible RAM limiting steps. The first is the storage of the subset of data, $\frac{nd}{s}$, the associated pairwise distance matrix $\left(\frac{n}{s}\right)^2$ and the compression set storage *sk* and its associated pairwise distance matrix $(sk)^2$. The most efficient possibility will be when these are of equal size and so;

(8)

$$\frac{nd}{s} + \left(\frac{n}{s}\right)^2 + sk = sk + (sk)^2$$

Which rearranges to

(9)

$$s^4k^2snd + n^2 = 0$$

Quartic equations such as this can be solved using Ferrari's solution, and since the terms *b* and *c* (s^3 and s^2) are both zero, and *a*, *d* and e (s^4, s^1, and s^0) are all positive, this will result in one real, positive solution for the most efficient number of subsets *s.*[2] If, as is the case for large MSI datasets, *n>>d* then only the pairwise distance matrices of either subsets or the compression set need be considered, resulting in a more general RAM requirement of the algorithm to be the pairwise distances of either a single subset of data or the compression set $\left(\frac{n}{s}\right)^2$ or $(sk)^2$. The minima of this will be when these are of equal size so $\left(\frac{n}{s}\right)^2 = (sk)^2$ or as with the two phase k-means, $\frac{n}{s} = sk$. As with the two phase *k*-means this gives the optimal number of subsets as $s = \sqrt{\frac{n}{k}}$ and so

(10)

$$RAM_{min} = \left(\frac{n}{\sqrt{\frac{n}{k}}}\right)^2$$

This rearranges to

(11)

$$RAM_{min} = nk$$

Therefore the space complexity of the two phase graph cuts approximates to O(n.k)

input : data matrix M of n pixels by d mass channels
input : k nearest neighbours
input : weighting function for graph generation
input : Number of eigenvector to cluster on e
input : k-means clustering parameters
output: Vector length n of cluster assignment

Construct a weighted graph W where $W_{i,j}$ represents the similarity between spectra i and j;
Perform eigendecomposition of W to get eigenvectors V and eigenvalues v;
Sort V in ascending order of the corresponding eigenvalues;
Perform k-means clustering on the V_{2-e+1};

Algorithm S1. Spectral clustering algorithm

input : data matrix of n pixels by d mass channels
input : desired number of subsets s
input : K-means clustering parameters
output: Vector length n of cluster assignment

Randomly assign the data into s equal sized subsets S_{1-s};
for $i \leftarrow 1$ **to** s **do**
    load into random-access memory (RAM) data assigned to subset S_i;
    cluster S_i into k clusters using the k-means algorithm;
    if $i == 1$ **then**
        create a compression set, a compressed representation of subset S_i from the cluster centroids;
    else
        add clutser centroids of subset S_i to the compression set;
    end
end
cluster the compression set using the k-means algorithm;
assign the cluster identities from the compression set to the pixels from subsets

Algorithm S2. Two-phase k-means clustering algorithm

input : data matrix of n pixels by d mass channels
input : desired number of subsets s
input : K-means clustering parameters
output: Vector length n of cluster assignment

Randomly assign the data into s equal sized subsets S_{1-s};
for $i \leftarrow 1$ **to** s **do**
    load into RAM data assigned to subset S_i;
    cluster S_i into k clusters using the spectral clustering algorithm;
    if $i == 1$ **then**
        create a compression set, a compressed representation of subset S_i from the cluster centroids;
    else
        add clutser centroids of subset S_i to the compression set;
    end
end
cluster the compression set using the spectral clustering algorithm;
assign the cluster identities from the compression set to the pixels from subsets

Algorithm S3. Two-phase graph cuts clustering algorithm

Figure S1. Comparison of *k*-means vs. graph cuts clustering on the brain datasets from figure 1 with different values of k (5, 10 and 15).

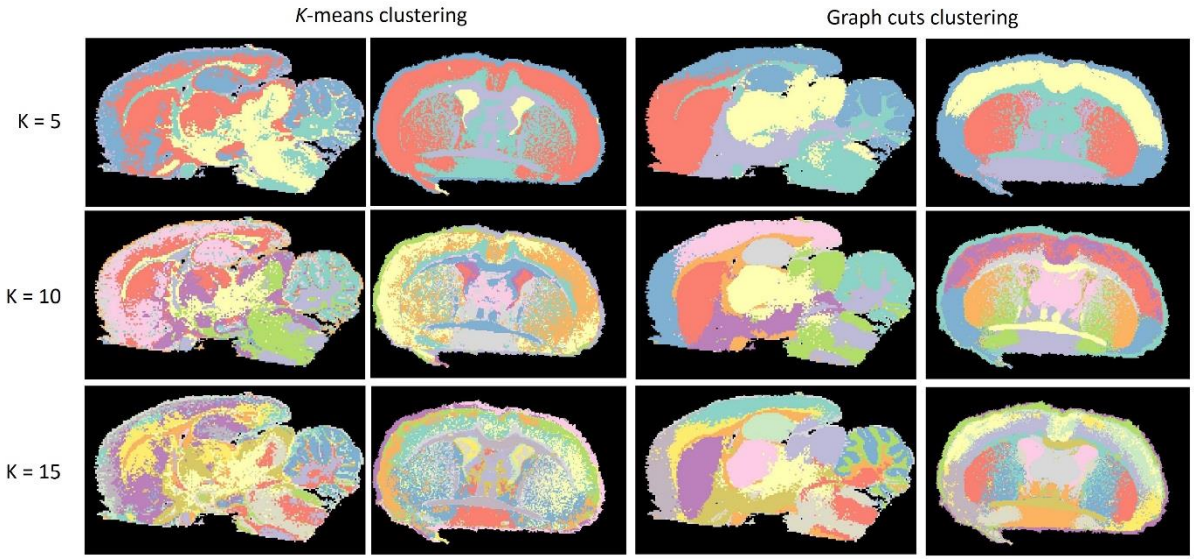


Figure S2. Example spectrum from coronal mouse brain acquired at 35.6 J m^{-2}.

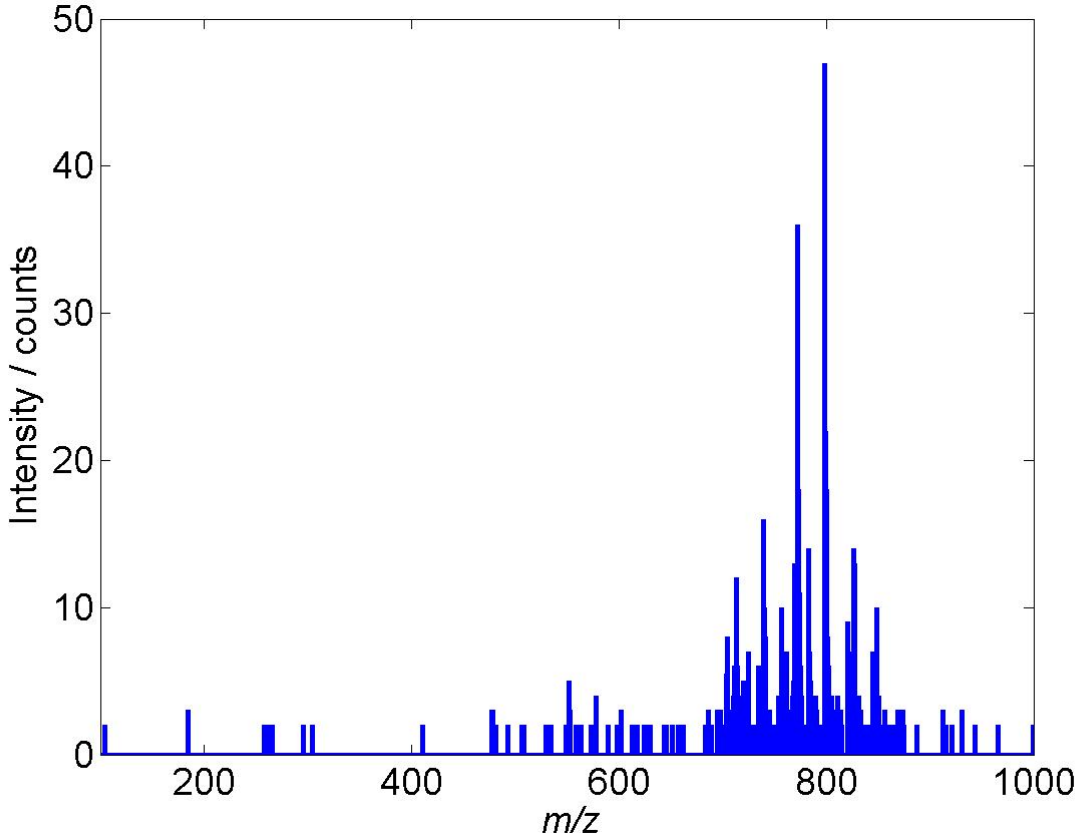


Figure S3. Example spectrum from coronal mouse brain acquired at 51.3 J m^{-2}.

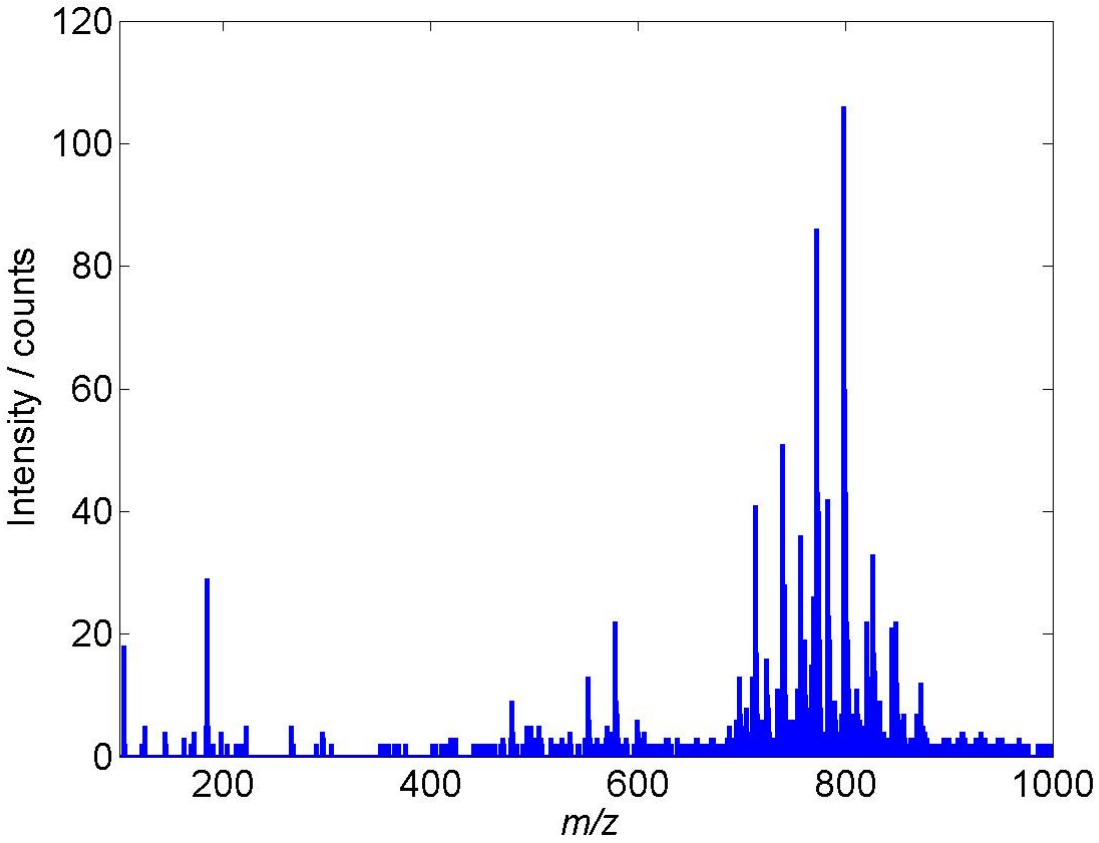


Figure S4. Example spectrum from coronal mouse brain acquired at 78.7 J m^{-2}.

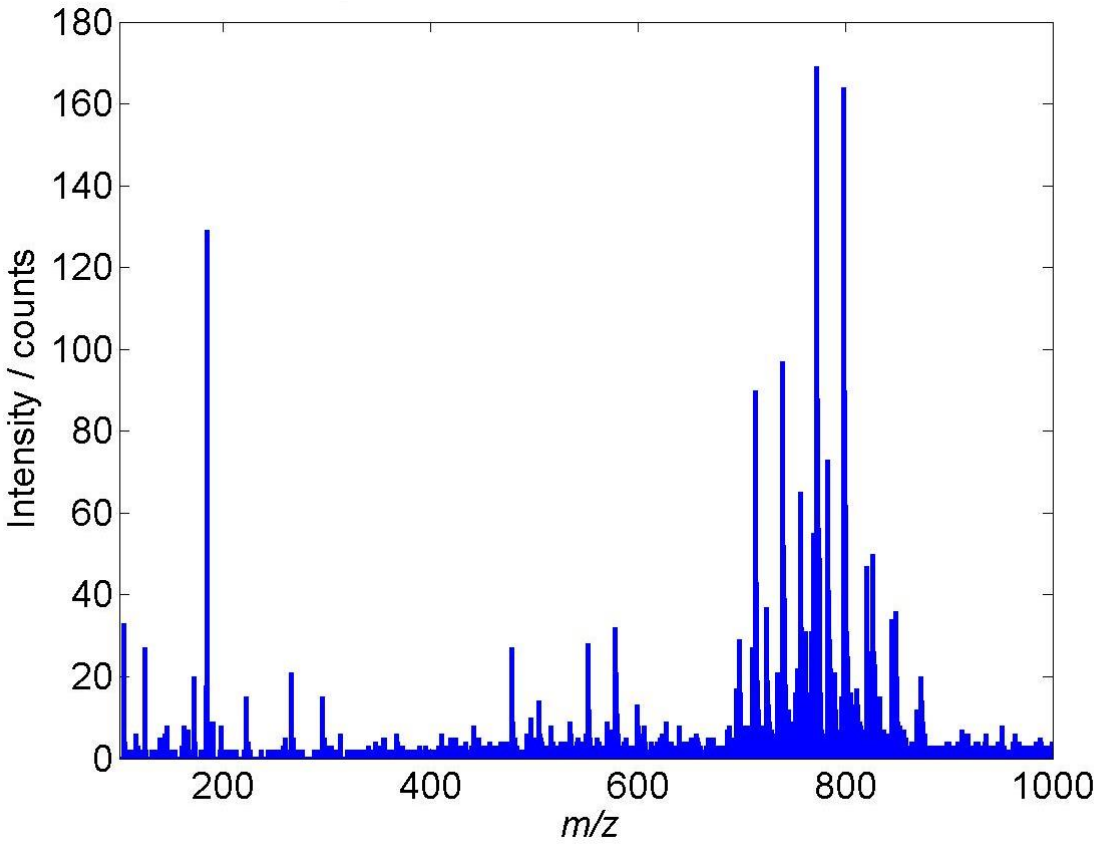


Figure S5. Example spectrum from coronal mouse brain acquired at 114.5 J m^{-2}.

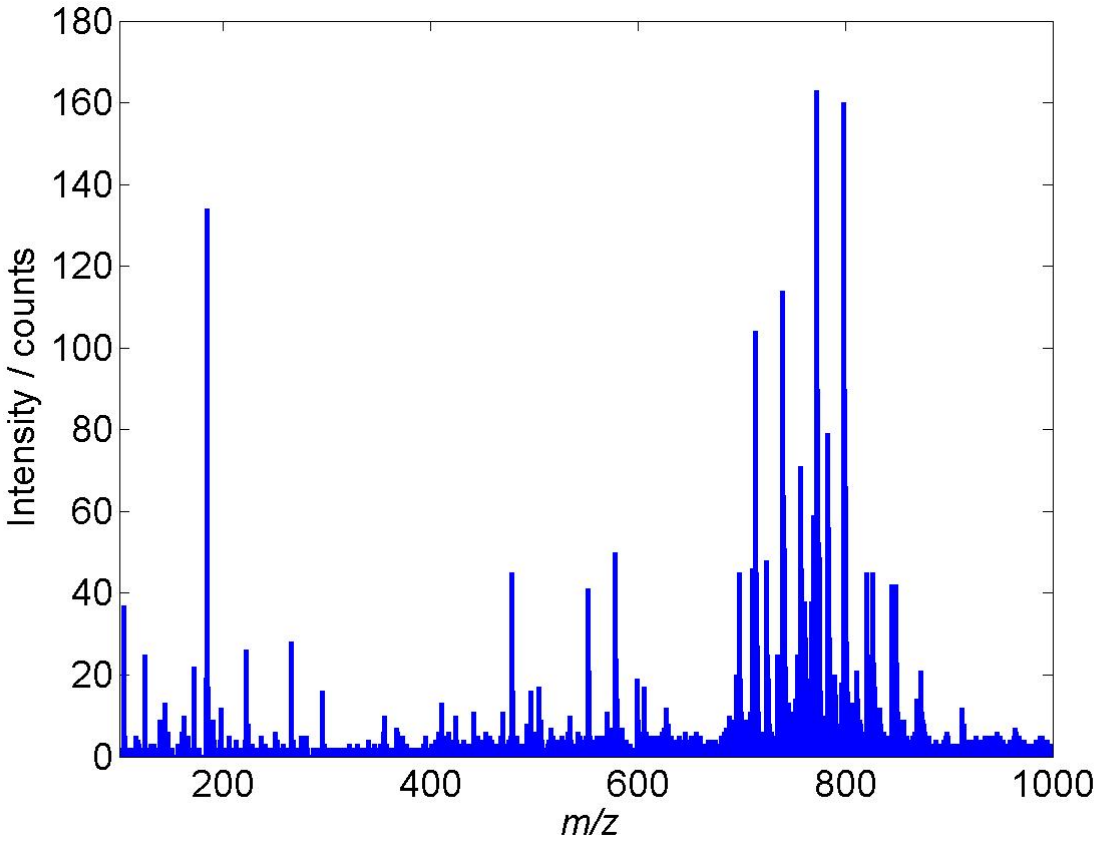


Figure S6. Example spectrum from coronal mouse brain acquired at 149.8 J m^{-2}.

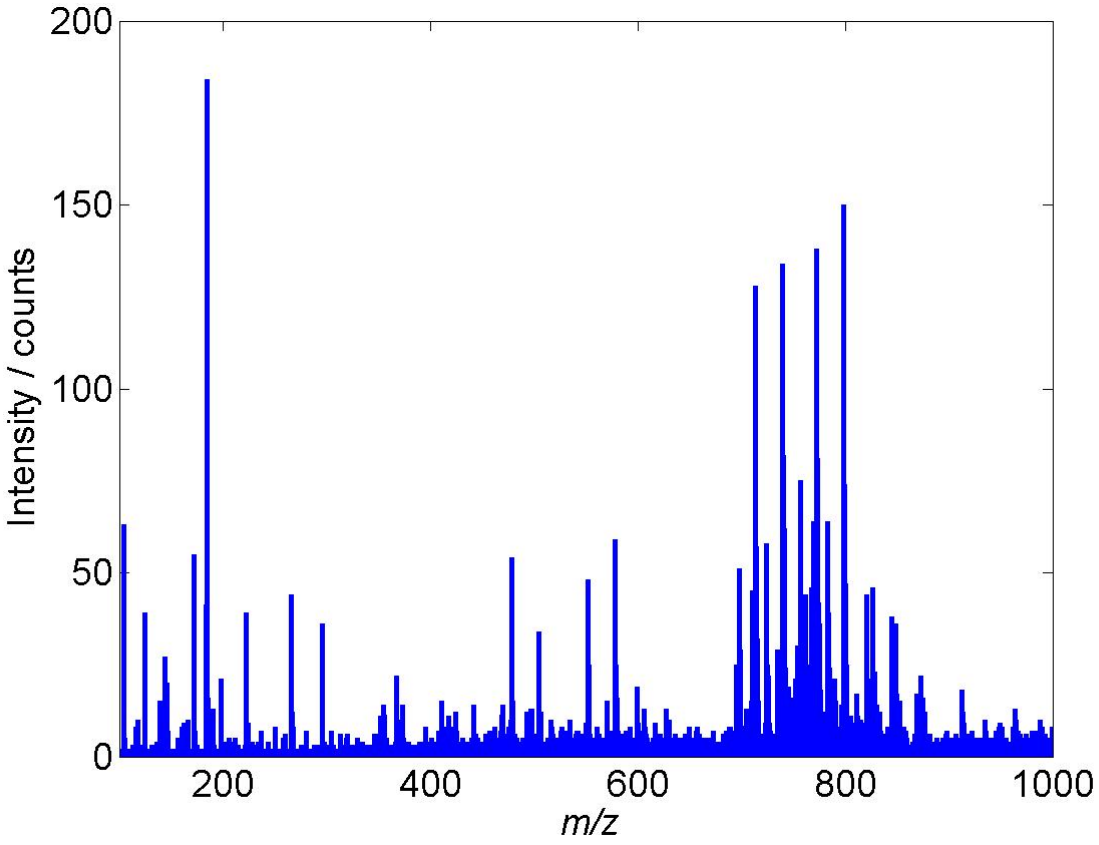


Figure S7. Result of graph cuts clustering on only the variable and control tissues acquired at 51.3 and 149.8 J m^{-2}.

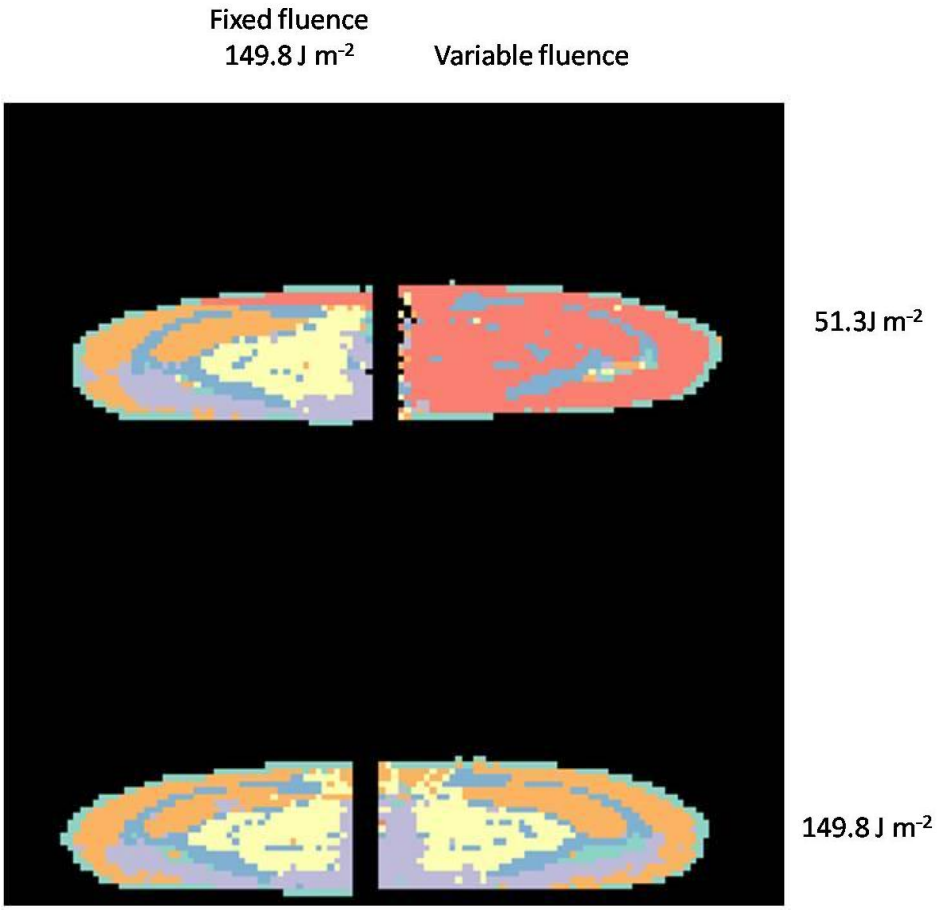


Figure S8. Time taken to perform *k*-means clustering and two-phase *k*-means clustering on synthetic data of varying sizes

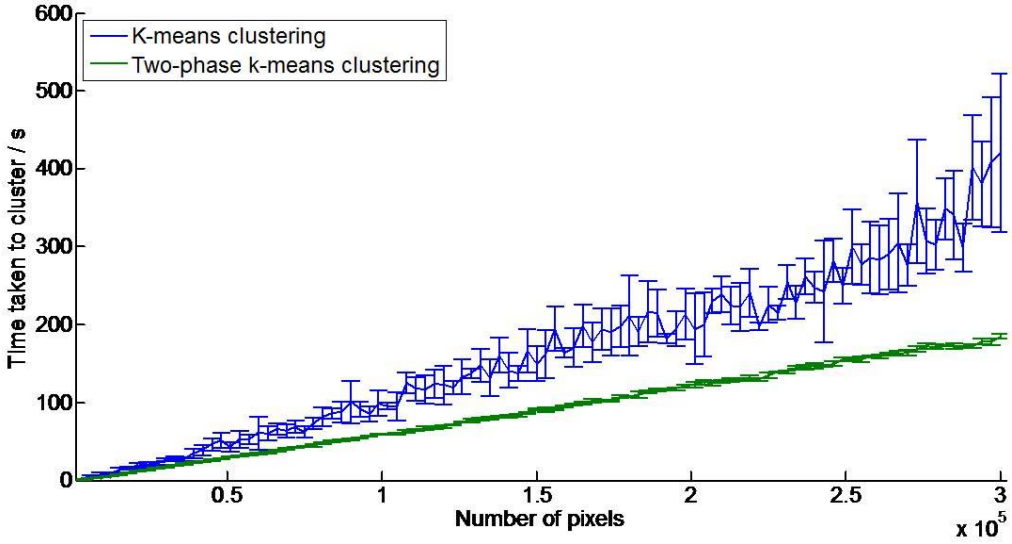


Figure S9. Sensitivity of *k*-means clustering and two-phase *k*-means clustering on synthetic data of varying sizes. Errorbars represent one standard deviation from the mean.

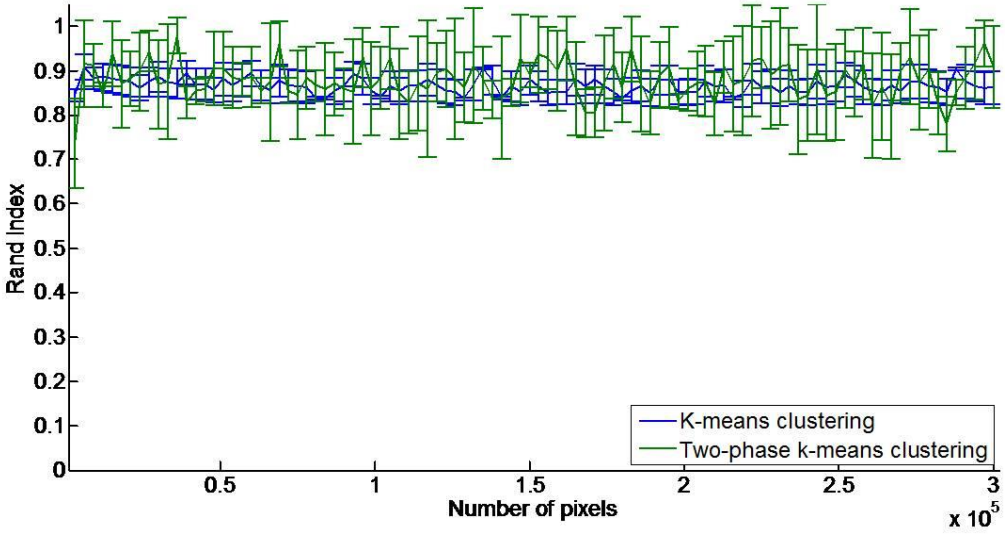


Figure S10. Number of peaks that can be retained vs. number of pixels in the image when loading a dataset into RAM.

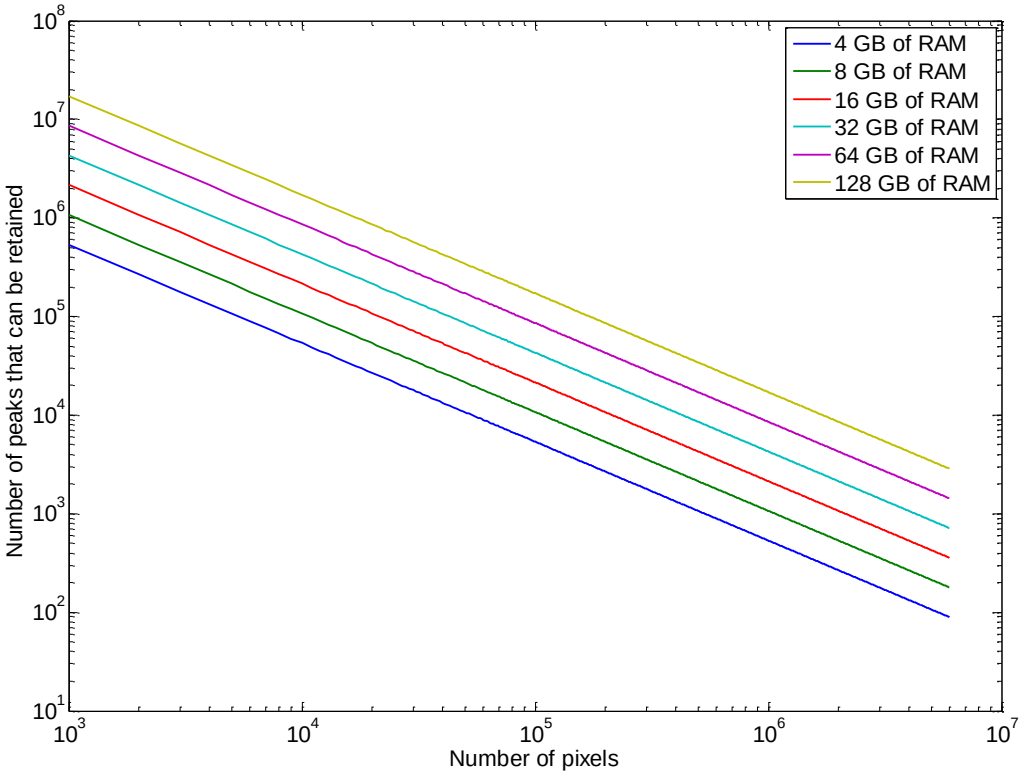


Figure S11. Number of peaks that can be retained vs. number of pixels in the image when loading subsets of the data into RAM using the two-phase clustering methods.

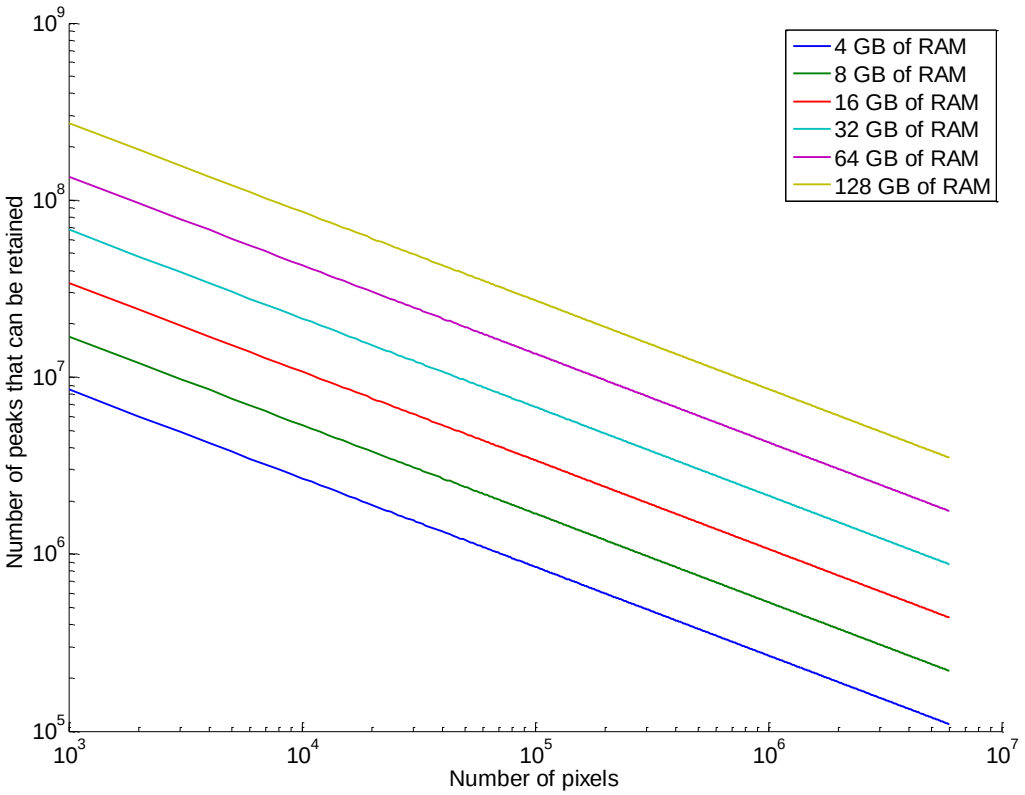
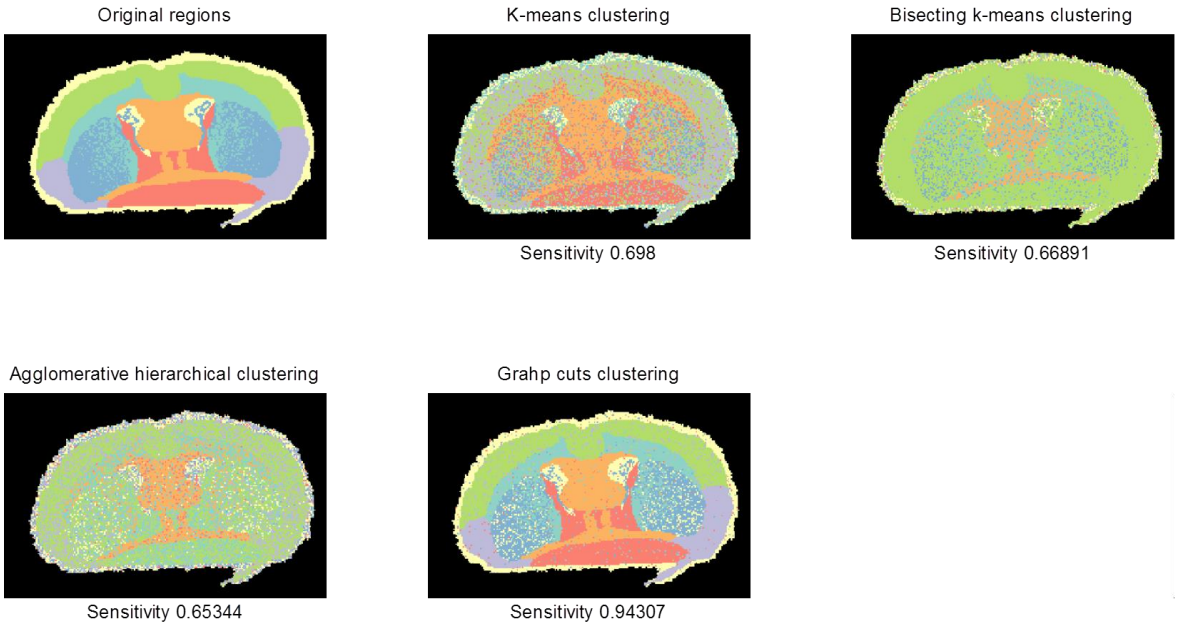


Figure S12. Comparison of clustering algorithms on a small synthetic dataset comprising of seven regions and 20,825 pixels, generated via statistical modelling.



References

(1) Shariatgorji, M.; Nilsson, A.; Goodwin, R. J.; Källback, P.; Schintu, N.; Zhang, X.; Crossman, A. R.; Bezard, E.; Svenningsson, P.; Andren, P. E. *Neuron* **2014**, *84*, 697-707.
(2) Neumark, S. *Solution of cubic and quartic equations*; Elsevier, 2014.