

# *massPix*: an R package for annotation and interpretation of mass spectrometry imaging data for lipidomics

Nicholas J. Bond<sup>1</sup> · Albert Koulman<sup>1</sup> · Julian L. Griffin<sup>1,2</sup> · Zoe Hall<sup>1,2</sup> 

Received: 18 May 2017 / Accepted: 14 August 2017 / Published online: 21 September 2017  
© The Author(s) 2017. This article is an open access publication

## Abstract

**Introduction** Mass spectrometry imaging (MSI) experiments result in complex multi-dimensional datasets, which require specialist data analysis tools.

**Objectives** We have developed *massPix*—an R package for analysing and interpreting data from MSI of lipids in tissue.

**Methods** *massPix* produces single ion images, performs multivariate statistics and provides putative lipid annotations based on accurate mass matching against generated lipid libraries.

**Results** Classification of tissue regions with high spectral similarity can be carried out by principal components analysis (PCA) or *k*-means clustering.

**Conclusion** *massPix* is an open-source tool for the analysis and statistical interpretation of MSI data, and is particularly useful for lipidomics applications.

**Keywords** Mass spectrometry imaging · Lipidomics · Bioinformatics software · Data processing

## 1 Introduction

Mass spectrometry imaging (MSI) is a transformative technology in systems biology and clinical research (Addie et al. 2015; Angel and Caprioli 2013). MSI enables the in situ analysis of tissue molecular composition for hundreds of metabolites and lipids simultaneously. Sophisticated approaches and software are therefore required in order to analyse and interpret the vast amount of data collected with each imaging experiment. As such, new bioinformatics tools and resources are needed to recreate molecular maps across tissue and probe statistical differences across a tissue slice using advanced pattern recognition tools, particularly in studies where disease processes need to be examined on a spatial basis (Alexandrov et al. 2010; Smentkowski et al. 2007; Van de Plas et al. 2007).

There have been various software packages released to view and analyse MSI data (Bemis et al. 2015; Gibb and Strimmer 2012; Källback et al. 2016; Parry et al. 2013; Verbeeck et al. 2014). Many tools including Biomap, DataCube Explorer, msIQuant and MSiReader do not perform multivariate statistical analysis, whilst others are vendor specific, e.g., ImageQuest (Thermo Scientific). Omnispect and Cardinal are freely available and perform multivariate analysis on data using non-negative matrix factorization and spatially-aware clustering approaches, respectively. However these software packages do not provide lipid feature annotation. Recently, a framework for false-discovery rate-controlled metabolite annotation for MSI has been developed as part of the METASPACE consortium, with great potential for stream-lining MSI data analysis (Palmer et al. 2017).

Here, we have developed *massPix*, an R-based package which processes MSI data, plots single ion distributions and performs multivariate statistics [principal components analysis (PCA) and clustering]. This software is different from

---

Availability and implementation: The source code, R package, documentation and test data are freely available to download from <https://github.com/hallz/massPix>. *massPix* is distributed under the GNU General Public Licence (version 3).

---

✉ Julian L. Griffin  
jlg40@cam.ac.uk

✉ Zoe Hall  
zlh22@cam.ac.uk

<sup>1</sup> MRC Human Nutrition Research, 120 Fulbourn Road, Cambridge CB1 9NL, UK

<sup>2</sup> Department of Biochemistry and Cambridge Systems Biology Centre, University of Cambridge, 80 Tennis Court Road, Cambridge CB2 1GA, UK

available tools, in that it has been designed specifically for lipidomics applications, enabling putative lipid annotations based on accurate mass. In addition, PCA and clustering may be performed to classify regions across tissue based on their lipid profiles (Hall et al. 2016, 2017). Furthermore the software is freely available, easy to implement by novices to R, and adaptable if required, by advanced users.

## 2 Implementation

*massPix* supports data in imzML format (Race et al. 2012; Schramm et al. 2012). Free converters for raw data to imzML are available from <http://www.imzML.org>. Whilst *massPix* has been developed for high resolution matrix assisted laser desorption ionisation (MALDI) data acquired with Thermo Scientific instrumentation, the software is vendor agnostic and can be applied to any data in imzML format independent of mass spectrometry platform. *massPix* is compatible with Windows, Mac and Linux operating systems, and requires at least sufficient RAM to load the entire experimental dataset into memory (for instance to process 3 GB image file, ~3.2 GB memory is used). *massPix* is run from the R scripting interface, however a detailed knowledge of R is not required to install and use the software. Those with advanced knowledge of R programming can adapt the source code for

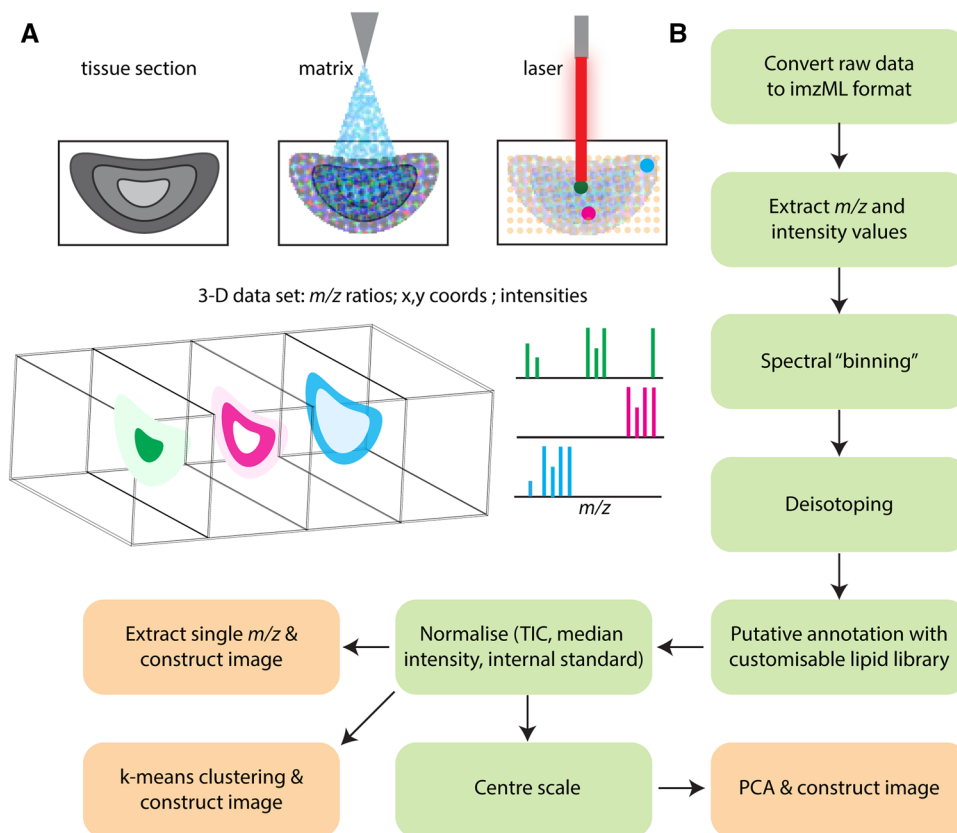
their own needs. *massPix* outputs high quality images, a data frame of the final normalised and annotated image which can be further manipulated in R, and csv files for spectra corresponding to cluster centers, PCA loadings, and lipid annotations. The *massPix* R package, all R scripts, library files and the imzML Converter are available on GitHub (<https://github.com/hallz/massPix>). A brief introduction is provided with parameter descriptions, in addition to a step-by-step presentation on software use and instructions on file conversion. Test data is available on the MetaboLights data repository (study ID: MTBLS487).

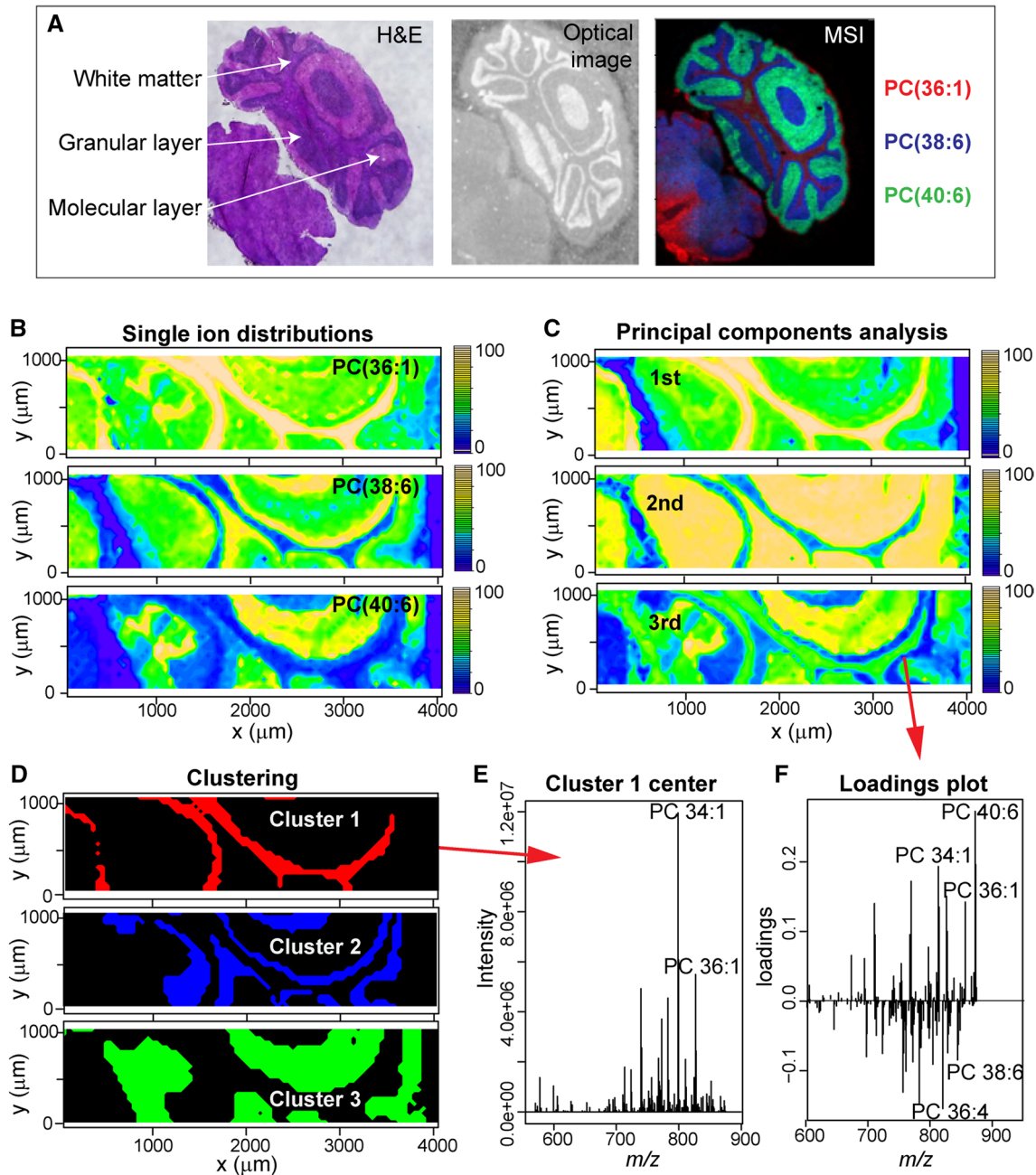
## 3 Results and discussion

### 3.1 Data acquisition

Most MSI workflows are based on MALDI or desorption electrospray ionisation (DESI) datasets. MALDI-MSI is currently more widely used within the field and these datasets have been used to develop *massPix*. In MALDI, a matrix is first applied to the tissue surface to aid ionisation. This is typically a small organic molecule, capable of absorbing the wavelength supplied by the laser and subsequently ionising surrounding analyte molecules (Fig. 1a). The laser raster-scans across the tissue surface, generating a

**Fig. 1** Overall MSI data acquisition and data analysis workflow. One of the most common mass spectrometry imaging approaches uses matrix assisted laser desorption ionisation (MALDI). First, a tissue section is coated with a matrix to aid ionisation. Then a laser is fired across the tissue, generating a spectrum ( $m/z$  ratios, ion intensities) for each pixel analysed ( $x, y$  coordinate) (a). The overall data processing workflow followed by *massPix* is shown (b)





**Fig. 2** Imaging mouse cerebellum using MALDI-MSI. H&E stained section of mouse cerebellum, with major tissue regions highlighted (*left*). An adjacent section was coated in matrix (*middle*) and analysed by MSI (*right*). Single ion distributions for  $[\text{PC}(36:1)+\text{K}]^+$ ,  $[\text{PC}(38:6)+\text{K}]^+$ ,  $[\text{PC}(40:6)+\text{K}]^+$ , shown in *red*, *blue* and *green*, are predominantly located in white matter, granular layer and molecular layer, respectively (**a**). Overlaid image produced using ImageQuest (Thermo Scientific). Single ion distributions produced by *massPix*

for  $[\text{PC}(36:1)+\text{K}]^+$ ,  $[\text{PC}(38:6)+\text{K}]^+$ ,  $[\text{PC}(40:6)+\text{K}]^+$ , in a sub-section of cerebellum (**b**). Principal components analysis (PCA) (**c**) and *k*-means clustering (**d**) differentiate regions based on their lipid profiles. Average spectra for pixels located in cluster 1 (**e**). PCA loadings plot for the third principal component (**f**); lipids with more positive (negative) loadings correspond to regions with higher (lower) principal component scores

mass spectrum for every pixel sampled. Spatial resolution is dependent on the optical design of the instrument, and varies from one to several hundred microns. The datasets generated are multi-dimensional, large and information-rich.

### 3.2 *massPix* pipeline

The overall data processing workflow (Fig. 1b) consists of initial data pre-processing, filtering, image subsetting,

deisotoping, annotation, normalisation, scaling, image “slicing” and multivariate statistics. First raw data must be converted to imzML format, which is then parsed to R. Ions with intensities greater than a threshold, from each spectra, are extracted and grouped to user-adjustable mass bins. The choice of bin width is dependent on the instrument mass resolving power (e.g. 10 ppm bin width for data acquired with 60,000 mass resolution at  $m/z$  400; for lower/higher resolving power increase or decrease bin width, respectively). Spectral features are defined by the median  $m/z$  value in each bin, and only features detected above a threshold proportion of spectra are retained. Average intensities for all features from a random subset of pixels are computed and used to perform deisotoping. The deisotoping algorithm identifies the molecular ion (M) and removes isotopes at  $m/z$  (M+1) and (M+2) which are within a calculated proportion of the intensity of M.

Putative lipid annotation by accurate mass is achieved by searching deisotoped ions against a generated library of lipid  $m/z$  ratios computed for all combinations of common fatty acids, lipid head-groups and anticipated adducts in each ionisation mode. The criteria for a match can be adjusted according to different MS performance capabilities (for example, <3, <10 ppm etc). Lipid classes searched in positive ion mode are diacylglycerides (DAG), triacylglycerides (TAG), phosphatidylcholines (PC), phosphatidylethanolamines (PE), phosphatidylserines (PS), LysoPC, cholesteryl esters (CE), sphingomyelins (SM) and ceramides (Cer). In negative ion mode, lipid classes searched are PC, phosphatidic acid (PA), PE, PS, phosphatidylglycerols (PG), phosphatidylinositols (PI), and free fatty acids (FFA). Whilst this list is not exhaustive, it does cover the most common lipid classes. Possible adducts considered are  $[M+K]^+$ ,  $[M+H]^+$ ,  $[M+Na]^+$ ,  $[M+NH_4]^+$  in positive ion mode and  $[M-H]^-$ ,  $[M+Cl]^-$ ,  $[M+OAc]^-$  in negative ion mode. It is important to point out that a database hit based on accurate mass should only be considered the first step in metabolite identification, and confirmation carried out using MS/MS is required, where this appropriate. This is particularly critical where data has been collected at lower mass accuracy, for instance using lower resolution time-of-flight instruments, where the risk of false positives is higher. For example, using the test data provided, an additional 200 possible lipid annotations were made by changing the mass accuracy for annotation from 5 to 50 ppm.

*massPix* has the further capability to perform difference matching on deisotoped features to search for mass differences associated with measurement-introduced alternation (e.g. fragmentation) or biological modifications (e.g. oxidation). Ion intensities are then normalised either to the median or total ion count, or to the average intensity of a set of standard ions. Single ion images can be produced, or normalised intensities used to create multivariate statistical

images based on  $k$ -means clustering or PCA following centering and Pareto scaling (van den Berg et al. 2006). The analysis can be readily customised by replacing default parameters for filtering, normalisation and scaling, library composition, lipid assignment and image reporting.

### 3.3 Test data

The power of multivariate statistics allows the differentiation of regions within tissue based on their lipid composition. This allows one to compare different regions in the same slice of tissue, for example tumour and adjacent tissue. As a test dataset, 15 micron tissue sections of wild type mouse cerebellum were coated with 2,5-dihydroxybenzoic acid (DHB) matrix (Sigma Aldrich, St Louis, MO; 10 mg/mL) and analysed by MSI (MALDI LTQ Orbitrap XL, Thermo Scientific, Hemel Hempstead, UK). The three major tissue regions within the cerebellum - white matter, granular and molecular layers (Fig. 2a)—were clearly differentiated by specific lipid profiles. Single ion distributions are shown for  $[PC(36:1)+K]^+$  (MSI Level 2; ChEBI:66857),  $[PC(38:6)+K]^+$  (MSI Level 2; ChEBI:64519),  $[PC(40:6)+K]^+$  (MSI Level 2; ChEBI:64431) which are predominantly located in white matter, granular layer and molecular layers, respectively (Fig. 2a, b). *massPix* uses an unsupervised approach to classify pixels of high spectral similarity using PCA (Fig. 2c) and  $k$ -means clustering (Fig. 2d). Spectra of cluster centres (Fig. 2e) and PCA loadings plots (Fig. 2f) provide detailed information about the relative lipid profiles of distinct regions and which lipid species are important for classification. The use of *massPix* software can thus aid interpretation of region-specific molecular changes. This is particularly important for understanding molecular mechanisms in disease processes.

**Acknowledgements** We thank Professor Timothy Cox and Dr Begona Cachon Gonzalez for providing samples of mouse brain, Dr Alan Race for advice on parsing imzML to R and Dr Sonia Liggi for software beta testing. This work was supported by the Medical Research Council (Lipid Profiling and Signalling [MC UP A90 1006] & Lipid Dynamics and Regulation [MC PC 13030]).

### Compliance with ethical standards

**Conflict of interest** The authors declare that they have no conflict of interest.

**Research involving animals** All applicable international, national, and institutional guidelines for the care and use of animals were followed.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give

appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

## References

- Addie, R. D., Balluff, B., Bovee, J. V., Morreau, H., & McDonnell, L. A. (2015). Current state and future challenges of mass spectrometry imaging for clinical research. *Analytical Chemistry*, 87(13), 6426–6433. doi:10.1021/acs.analchem.5b00416.
- Alexandrov, T., Becker, M., Deininger, S. O., Ernst, G., Wehder, L., Grasmair, M., et al. (2010). Spatial segmentation of imaging mass spectrometry data with edge-preserving image denoising and clustering. *Journal of Proteome Research*, 9(12), 6535–6546. doi:10.1021/pr100734z.
- Angel, P. M., & Caprioli, R. M. (2013). Matrix-assisted laser desorption/ionization imaging mass spectrometry: In situ molecular mapping. *Biochemistry*, 52(22), 3818–3828. doi:10.1021/bi301519p.
- Bemis, K. D., Harry, A., Eberlin, L. S., Ferreira, C., van de Ven, S. M., Mallick, P., et al. (2015). Cardinal: An R package for statistical analysis of mass spectrometry-based imaging experiments. *Bioinformatics*, 31(14), 2418–2420. doi:10.1093/bioinformatics/btv146.
- Gibb, S., & Strimmer, K. (2012). MALDIquant: A versatile R package for the analysis of mass spectrometry data. *Bioinformatics*, 28(17), 2270–2271. doi:10.1093/bioinformatics/bts447.
- Hall, Z., Ament, Z., Wilson, C. H., Burkhart, D. L., Ashmore, T., Koulman, A., et al. (2016). Myc expression drives aberrant lipid metabolism in lung cancer. *Cancer Research*, 76(16), 4608–4618. doi:10.1158/0008-5472.can-15-3403.
- Hall, Z., Bond, N., Ashmore, T., Sanders, F., Ament, Z., Wang, X., et al. (2017). Lipid zonation and phospholipid remodeling in non-alcoholic fatty liver disease. *Hepatology*, 65, 1165–1180.
- Källback, P., Nilsson, A., Shariatgorji, M., & André, P. E. (2016). msIQuant—quantitation software for mass spectrometry imaging enabling fast access, visualization, and analysis of large data sets. *Analytical Chemistry*, 88(8), 4346–4353. doi:10.1021/acs.analchem.5b04603.
- Palmer, A., Phapale, P., Chernyavsky, I., Lavigne, R., Fay, D., Tarasov, A., et al. (2017). FDR-controlled metabolite annotation for high-resolution imaging mass spectrometry. *Nature Methods*, 14(1), 57–60. doi:10.1038/nmeth.4072.
- Parry, R. M., Galhena, A. S., Gamage, C. M., Bennett, R. V., Wang, M. D., & Fernandez, F. M. (2013). omniSpect: An open MATLAB-based tool for visualization and analysis of matrix-assisted laser desorption/ionization and desorption electrospray ionization mass spectrometry images. *Journal of the American Society for Mass Spectrometry*, 24(4), 646–649. doi:10.1007/s13361-012-0572-y.
- Race, A. M., Styles, I. B., & Bunch, J. (2012). Inclusive sharing of mass spectrometry imaging data requires a converter for all. *Journal of Proteomics*, 75(16), 5111–5112. doi:10.1016/j.jprot.2012.05.035.
- Schramm, T., Hester, A., Klinkert, I., Both, J.-P., Heeren, R. M. A., Brunelle, A., et al. (2012). imzML—A common data format for the flexible exchange and processing of mass spectrometry imaging data. *Journal of Proteomics*, 75(16), 5106–5110. doi:10.1016/j.jprot.2012.07.026.
- Smentkowski, V. S., Ostrowski, S. G., Braunstein, E., Keenan, M. R., Ohlhausen, J. A., & Kotula, P. G. (2007). Multivariate statistical analysis of three-spatial-dimension TOF-SIMS raw data sets. *Analytical Chemistry*, 79(20), 7719–7726. doi:10.1021/ac071019o.
- Van de Plas, R., Ojeda, F., Dewil, M., Van Den Bosch, L., De Moor, B., & Waelkens, E. (2007). Prospective exploration of biochemical tissue composition via imaging mass spectrometry guided by principal component analysis. *Pacific Symposium on Biocomputing*, 12, 458–469.
- van den Berg, R. A., Hoefsloot, H. C., Westerhuis, J. A., Smilde, A. K., & van der Werf, M. J. (2006). Centering, scaling, and transformations: Improving the biological information content of metabolomics data. *BMC Genomics*, 7, 142. doi:10.1186/1471-2164-7-142.
- Verbeeck, N., Yang, J., De Moor, B., Caprioli, R. M., Waelkens, E., & Van de Plas, R. (2014). Automated anatomical interpretation of ion distributions in tissue: Linking imaging mass spectrometry to curated atlases. *Analytical Chemistry*, 86(18), 8974–8982. doi:10.1021/ac502838t.