Chapter 10
Statistical Procedure for IMS Data Analysis

Yuki Sugiura and Mitsutoshi Setou

Abstract In MALDI-IMS of tissue samples, since the tissue contains an enormous variety of biomolecules, a complex mass spectrum with hundreds of to a thousand peaks can be obtained from a single data point. Furthermore, several thousands of spectra with spatial data are obtained at one IMS experiment. Because of the complexity and enormousness of the IMS dataset, manual processing of the dataset to obtain significant information (e.g., identification of disease-specific mass signature) is not a realistic procedure. In this regard, today, multivariate analysis becomes a powerful tool in IMS data analysis. In this chapter, we describe an unsupervised multivariate data analysis technique that enables us to sort the data sets without any reference information. Particularly, two major methods that are related to IMS, namely, hierarchical clustering and principal component analysis (PCA), are described in detail with examples. Finally a basic procedure for PCA with familiar software (such as Microsoft Excel) is introduced.

10.1 Introduction

A single mass spectrum contains a lot of useful information. In an imaging mass spectrometry (IMS), the spectra are further added to spatial information by scanning the tissue sample two-dimensionally. Moreover, a typical IMS-mass spectrum contains hundreds to a thousand peaks because tissue sections contain an enormous number of biomolecules. Furthermore, the number of obtained spectra could be several thousands in one IMS measurement. The recent advances of IMS regarding

Y. Sugiura
Department of Bioscience and Biotechnology, Tokyo Institute of Technology,
4259 Nagatsuta-cho, Midori-ku, Yokohama, Kanagawa 226-8501, Japan

Y. Sugiura and M. Setou (✉)
Department of Molecular Anatomy, Hamamatsu University School of Medicine,
1-20-1 Handayama, Higashi-ku, Hamamatsu, Shizuoka 431-3192, Japan
e-mail: setou@hama-med.ac.jp

M. Setou (ed.), *Imaging Mass Spectrometry*: *Protocols for Mass Microscopy*,
DOI 10.1007/978-4-431-09425-8_10, © Springer 2010

simultaneous detection of numerous molecules (i.e., large number of mass peaks) at high spatial resolution (i.e., large number of data points) further enlarge the volume of IMS datasets. Because of such enormity of the IMS dataset, the development of statistical analyses is essential, and several multivariate analyses have been developed as useful tools for extracting important information. In particular, to date, "unsupervised" multivariate data analysis techniques have been widely utilized; they enable one to sort datasets without any reference information [1–4].

As Yanagisawa et al. reported in an early study [4], unsupervised multivariate analysis – particularly hierarchical clustering – helps one extract clinically important information for diagnosis/prognosis purposes, from a large-scale dataset. Also, principal component analysis (PCA) [1–3] and independent component analysis (ICA) [5] have been used to reduce data, thus enabling the extraction of the specific mass peaks [6, 7] and tissue locations of interest (e.g., normal tissue region vs. cancerous region) [7]. More recently, a more readily interpreted biological/clinical method using probabilistic latent semantic analysis (pLSA) has been developed [8].

In this chapter, we describe a quite basic protocol for an unsupervised multivariate data analysis technique, particularly PCA, using only Microsoft Excel and free software. Although several useful software packages commercially available provide much fast and easier data analysis with automatic calculation (just click "Run" command; see Chap. 11), the aim of this chapter is introduction of statistical analysis with minimal procedure for one who is not statistically expert. It is strongly recommended that each experimenter perform the analysis himself or herself using standard statistical tools such as Microsoft Excel; such experience should help the experimenters in later performing more complicated analyses governed by software packages that proceed automatically with calculation.

10.2 IMS Linked to Multivariate Analysis

Multivariate analysis is applied to a dataset that involves more than one statistical variable at a time. Because large numbers of variables are involved in matrix-assisted laser desorption/ionization (MALDI)-IMS – as a result of the mass spectrum combined with spatial information, researchers find multivariate analysis effective in extracting biological/clinical information. Hierarchical clustering is used to sort datasets into several clusters, according to similarities among variables. The traditional representation of this hierarchy is a tree (called a dendrogram), with individual elements at one end and a single cluster containing every element at the other. Agglomerative algorithms begin at the leaves of the tree, whereas divisive algorithms begin at the root [9].

In IMS (or direct tissue profiling with MS), each of the MS profiles could be clustered, based on the m/z value and its intensity, i.e., peak expression pattern. Employing this technique, one can identify the mass peaks that are different between control and diseased samples (or regions), thus possibly identifying

biomarkers for specific patients by the molecular signature. Yanagisawa et al. performed direct MS analysis of lung tissue obtained from non-small cell lung cancer (NSCLC) or nontumor patients and processed the direct tissue profiling data via the clustering [4]. As shown in Fig. 10.1 (top panel), the profiles of NSCLC and nontumor patients can be clearly distinguished. In addition, the primary alteration can be discriminated from others (bottom panel). Such an unsupervised clustering analysis enables not only sorting data into clusters but also extracting reliable markers according to the statistical criteria. In the technical aspect, in this example, by converting mass spectra into the data matrix of a peak intensity list, the authors reduced the volume of data, enabling an efficient analysis.

PCA, on the other hand, has been utilized to understand the overview of the spatial molecular distribution patterns [1–3]. We do not describe the detailed mathematical theory for reasons of space limitation, but in brief, it is a statistical method that merges the data containing multiple elements into low-dimensional data.

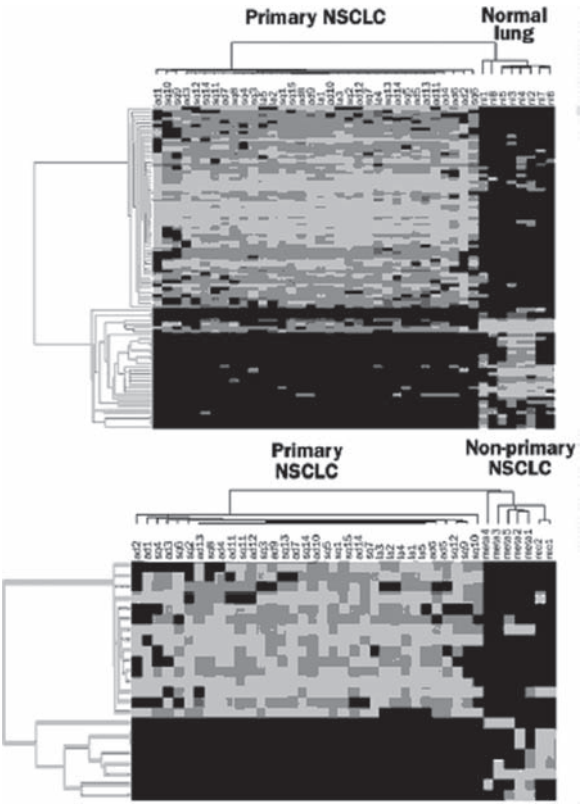


Fig. 10.1 An example of clustering according to peak intensity, calculated from mass spectra. Thirty-four non-small cell lung cancer (*NSCLC*) patients were clearly distinguished from eight controls. In addition, primary patients could be distinguished from nonprimary patients [4] (Reprinted from Yanagisawa et al., Lancet 362:433–439.)

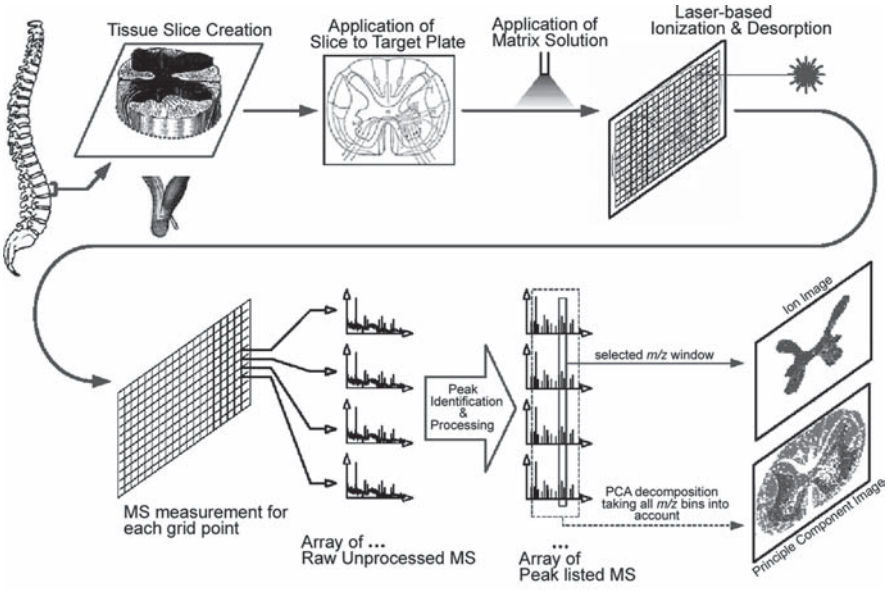


Fig. 10.2 After performing standard IMS using spinal cord sections, each peak was detected to obtain two-dimensional (2D) images of ions. At the same time, PCA analyses were performed by using intensity values derived from all the mass peaks [3] (Reprinted from Plas et al., 2007 IEEE/NIH Life Science Systems and Applications Workshop, pp 209–212.)

It reduces a large set of variables to a small set of variables called "principal factors," which are linear combinations of the original variables.

Figure 10.2 shows an example of PCA-coupled IMS data analysis. In the strategy reported by Plas et al., spectra obtained by IMS are processed to peak detection, and based on the generated peak list, two-dimensional (2D) ion images were reconstructed and PCA decomposition was performed. PCA images (i.e., 2D ion intensity map of principal component score on the tissue section) were utilized to find trends of proteomic distribution patterns. By utilizing IMS-PCA on spinal cord tissue section, Plas et al. demonstrated that proteomic composition in the "butterfly-shaped" posterior column of the spinal cord differs considerably from that in other regions of the spinal cord, according to statistical criteria [3].

Figure 10.3 shows another example, in which the IMS-PCA of a breast cancer section was performed. In this case, PCA revealed that the largest spectral differences (i.e., the largest difference in proteomic composition) were observed between connective tissue and cancer area (in the principal component 1 image, panel e). Furthermore, the second largest differences in protein expression pattern were observed within two tumor cell populations, which are HER2-positive/negative cells (panel b), revealed by the principal component 2 image (panel f) [10].

PCA is also helpful in identifying the meaning variables, for example, identifying which mass peaks are responsible for making difference between normal and drug-treated tissue regions. Such mass peaks can be candidates for the bio-

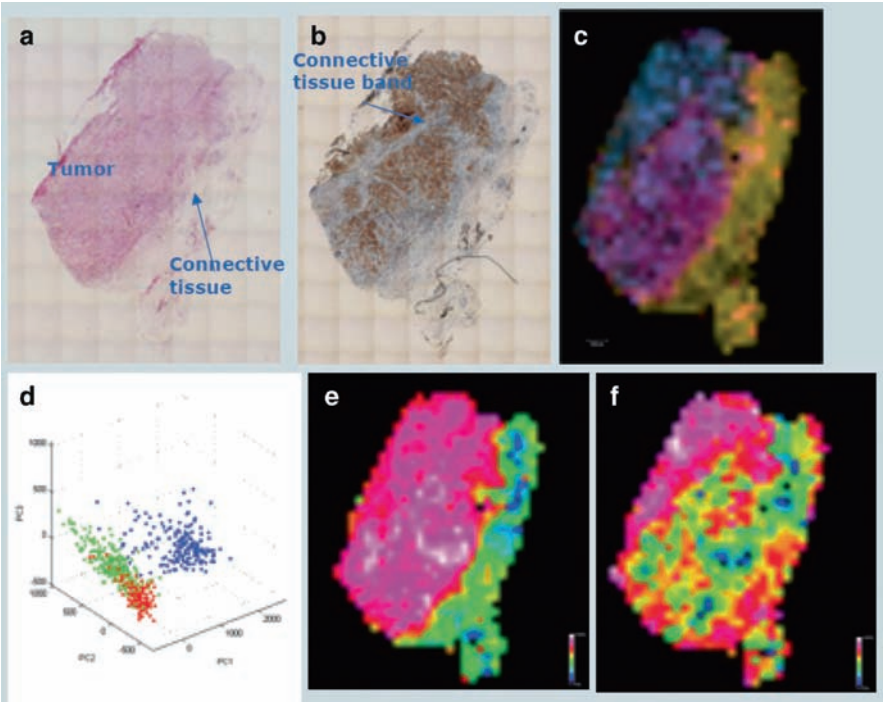


Fig. 10.3 IMS linked PCA with breast cancer section. Optical images of hematoxylin and eosin (H&E)-stained (**a**) and HER2-immunostained (**b**) breast cancer sections. **c** Ion image at several mass peaks were stained with *blue* and *red*. **d** Graph in which principal component scores for PC1, PC2, and PC3 are plotted. **e** Principal component 1 image. According to the value of principal component score calculated for each spectrum at each tissue location, pixels were stained with pseudo-color. **f** Principal component 2 image (Kindly permitted by PD Dr. med. Axel Walch, Institute for Pathology, Helmholtz Center Munich, German Research Center for Environmental Health (GmbH) Ingolstadter Landstrasse 1, 85764 Oberschleissheim, Germany)

marker [11, 12]. Prideaux et al. showed the altered metabolic profiling of porcine skin after treatment with a commercial 0.1% hydrocortisone cream. By utilizing PCA, they successfully sorted the two groups of skin MS profiles with/without drug treatment and identified the altered peak expressions (Fig. 10.4) [6]. Such direct MS profile coupling PCA can be applied to analysis in micro-tissue domains as small as a laser spot (diameter >100 μm) that are difficult to separate and analyze by using conventional approaches [6, 7].

10.3 IMS-PCA on the Genetically Manipulated Mouse Brain

In the following sections, we introduce a simple protocol to identify mass peaks of which difference between samples by PCA. We used only Microsoft Excel and SpecAlign software, the latter of which is a graphic computational tool for

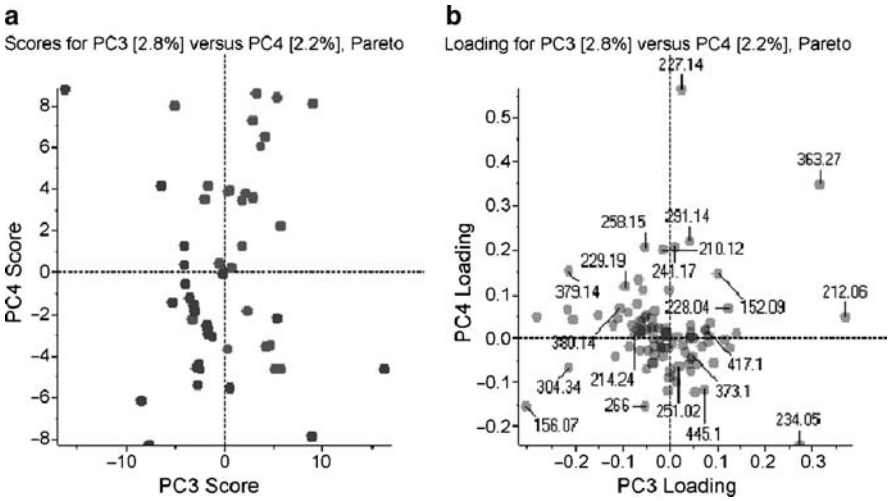


Fig. 10.4 **a** Principal component analysis of 25 spectra taken from inside and 25 spectra from outside the area of porcine skin treated with a commercial 0.1% hydrocortisone cream. In the score plot, a clear grouping of treated (*red spots*) and untreated (*blue squares*) spectra is observed. **b** The factor loading plot indicates that *m/z* 363.28, the $[M + H]^+$ ion for hydrocortisone, is a major contributor to differentiation among the groups (Reprinted from Prideaux et al., Int J Mass Spectrom 260:243–251.)

spectrum processing. (SpecAlign is free of charge and available for download from Wong et al., from http://ptcl.chem.ox.ac.uk/jwong/specalign [13].)

Here, we analyzed animals with distinct genetic backgrounds; they were either wild type (WT) or SCRAPPER (Scr) knockout (KO) mice lacking a gene coding for a ubiquitin ligase [14]. We applied IMS-linked PCA, to compare the proteomic composition of WT and Scr-KO mice brains and further searched for substances differentiating the two genotypes.

10.4 IMS of WT and Scr-KO Mouse Brain Sections

In Scr-KO mice, obvious pathological features were observed, particularly in their brain. For example (Fig. 10.5), hematoxylin and eosin (H&E) staining of Scr-KO brain tissue revealed the existence of sponge-like degeneration. It has been reported that the mutant mice of other ubiquitin E3 ligase exhibit an age-dependent neuropathology, including spongiform degeneration [15]. The release of neurotransmitter in Scr-KO mice is abnormal [14]; such neural dysfunction possibly induced the histological degeneration observed in Fig. 10.5.

There are several ways to determine the molecules that are involved in such histological degeneration. For example, in immunohistochemistry and immunoblotting, specific alterations of proteins with high sensitivity and specificity can be revealed; however, analyses require specific antibodies to the candidate protein.

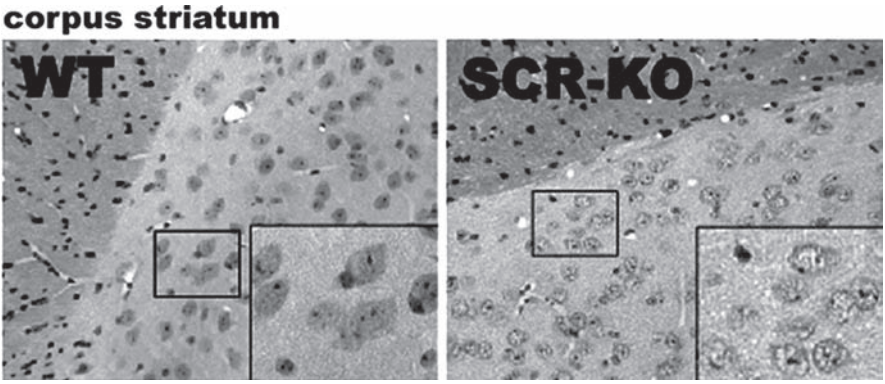


Fig. 10.5 Optical image of hematoxylin and eosin (H&E)-stained corpus striatum region of wild-type (*WT*) and Scr-knockout (*SCR-KO*) mice [7] (Reprinted from Yao et al., Proteomics 8:3692–3701.)

Because generation of the antibodies requires large cost and much time, one must determine the small number of candidate molecules that might be involved in the defect before the experiment.

In this regard, MS is able to detect molecular alterations without any tags or probes, even of unexpected molecules. In this context, MS-based imaging can identify such differences with spatial information, even within micro-tissue regions that are difficult to dissect out.

10.4.1 Materials Used and Measurement Conditions

The following are the details of an IMS experiment using sagittal brain sections obtained from WT and Scr-KO mice:

- Brain sections, sliced at 5-μm thickness [16], mounted on ITO glass slide (Bruker Daltonics)
- 70% Ethanol for rinsing tissue sections, 30 s × two times [17]
- Matrix: Sinapic acid (Bruker Daltonics) [25 mg ml^{-1} in 0.1% trifluoroacetic acid (TFA), 50% acetonitrile (v/v)], applied to section by spray-coating method
- Mass spectrometer: MALDI-TOF/TOF-type instrument (Ultra Flex 2, Bruker Daltonics)
- Measurement modes: positive-ion detection mode, linear mode
- Laser interval: 80 μm
- Software used for image reconstitution and spectrum-extraction from region of interest: flexImaging 2.0 (Bruker Daltonics)

Acquired IMS data were sorted and preprocessed before PCA, as follows:

- Following IMS measurement, mass spectra were collected from each brain region of WT and Scr-KO mice (using flexImaging 2.0 software); in this study,

those regions were composed of the cerebral cortex, pons, hypothalamus, and corpus striatum (see Fig. 10.13 later in this chapter).
- The spectral data were then formatted into ASCII code, using flexAnalysis 3.0 software, to allow processing by SpecAlign software.
- Next, the converted spectra were normalized to equalize total ion current by using SpecAlign software [13]. This is an important process, as it rescales the sample-to-sample variability of the peak intensity values before proceeding to statistical analysis [6, 18, 19].

In data analysis procedures without statistical methods, we usually averaged the spectra of each region and visually compared the mass peaks between samples one by one. With such visual comparisons of spectra (Fig. 10.6), we were certainly able to find differences among the peak expressions, as indicated by the arrows; this, in turn, indicates the abnormal expression/suppression of proteins in KO mice. However, such methodology is inefficient, especially when one is analyzing a large number of regions and/or many tissue sections.

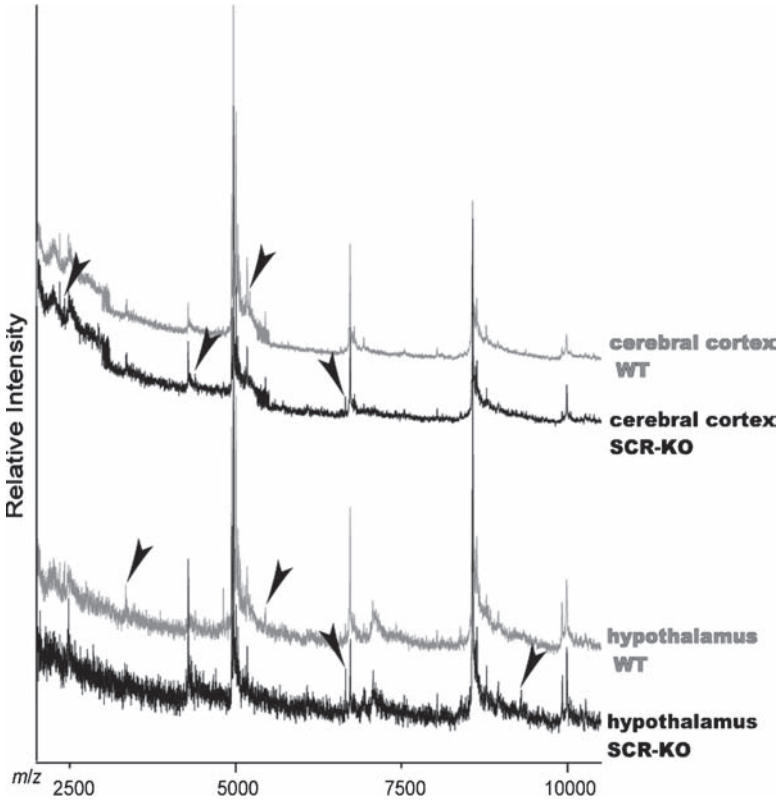


Fig. 10.6 Comparison of normalized spectra following extractions from *WT* and *SCR-KO* mice [7] (Reprinted from Yao et al., Proteomics 8:3692–3701.)

10.4.2 IMS-PCA of WT and Scr-KO Mouse Brains

Below, we describe IMS-linked PCA to compare the proteomic composition of the corpus striatum region, in which severe pathological alterations were found in Scr-KO mice (see Fig. 10.5).

For PCA, to reduce the amount of calculation required, we generated mass peak lists from the sets of extracted spectra rather than use of every continuous *m*/*z* value; For this operation, peak detection function of SpecAlign software was used [13]. Although such a scenario would be ideal, using every *m*/*z* value of every spectrum as a variable would claim too many calculations. Figure 10.7 shows the peak list file used for PCA, created by importing peak list files into Microsoft Excel [arranged as *x*-axis = *m*/*z* (*validate*), *y*-axis = spectrum (*case*)].

Next, PCA was performed by Microsoft Excel add-in software. Several free or low-cost software packages that can be used as add-in or macro tools for Microsoft Excel are available. Generally, we must define the "case" and "variable" for PCA calculation; in this case, we defined each mass spectrum with spatial information as the *case*, and each includes approximately 80 distinct mass peak intensities (*variables*).

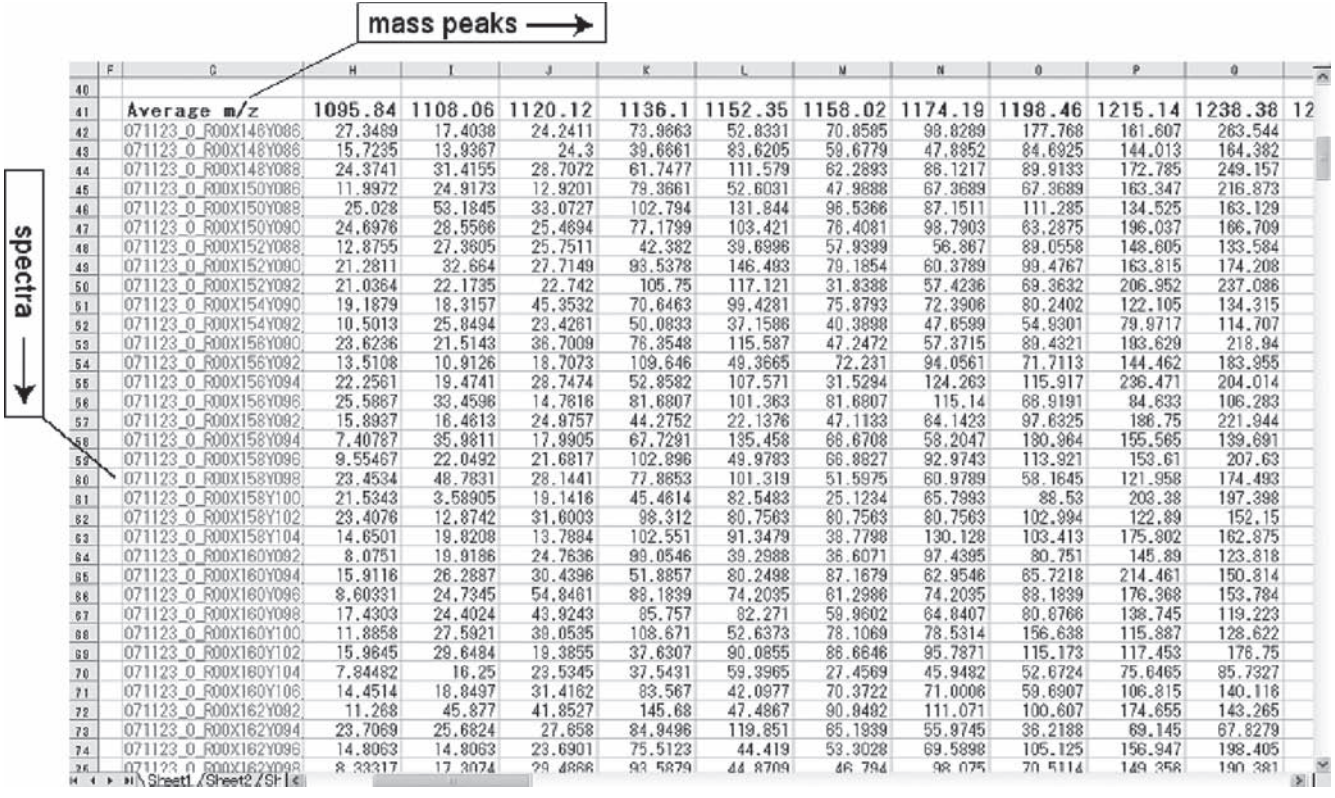
10.4.3 Data Interpretation of PCA

10.4.3.1 Interpretation of Component Scores

As a result of undertaking PCA, several parameters can be obtained. Among those, the *component score* and *factor loading* are important postanalysis steps. A component score is calculated for each mass spectrum; all are defined for each principal component (e.g., for PC1, PC2.…) (Fig. 10.8).

Those component scores can be plotted two-dimensionally to facilitate interpretation of the PCA results. In Figs. 10.9 and 10.10, component scores for PC1 and PC2 are plotted on the *x*-axis and *y*-axis, respectively. and each dot represents a spectrum from a distinct tissue sample or location. What is important to note is whether two populations of spectra (= dot) obtained from distinct samples (e.g., normal vs. diseased) are "spatially separated" on the graph. If they are separated (see Fig. 10.9a), it means that the molecular expression patterns of these two regions were statistically different from each other. If not, PCA failed to extract the statistical differences between the two populations (see Fig. 10.9b).

In Fig. 10.10, open and filled spots indicate spectra obtained from WT and Scr-KO mice, respectively. Notably, two populations of spots are spatially separated on this graph (as represented by open and filled ellipses), indicating that those two sets of spectra from WT and Scr-KO mice tissue contain proteomic differences. We focused on the *y*-axis (PC2) that separates nicely between the two populations: one could find that the spectra of WT mice had positive component scores against PC2 whereas Scr-KO mice had negative scores.



	C	H	I	J	K	L	M	N	O	P	Q
41	Average m/z	1095.84	1108.06	1120.12	1136.1	1152.35	1158.02	1174.19	1198.46	1215.14	1238.38
42	071123_0_R00X146Y086	27.3489	17.4038	24.2411	73.9663	52.8331	70.8585	98.8289	177.768	161.607	263.544
43	071123_0_R00X148Y086	15.7235	13.9367	24.3	39.6661	83.6205	59.6779	47.8852	84.6925	144.013	164.382
44	071123_0_R00X148Y088	24.3741	31.4155	28.7072	61.7477	111.579	62.2893	86.1217	89.9133	172.785	249.157
45	071123_0_R00X150Y086	11.9972	24.9173	12.9201	79.3661	52.6031	47.9888	67.3689	67.3689	163.347	216.873
46	071123_0_R00X150Y088	25.028	53.1845	33.0727	102.794	131.844	96.5366	87.1511	111.285	134.525	163.129
47	071123_0_R00X150Y090	24.6976	28.5566	25.4694	77.1799	103.421	76.4081	98.7903	63.2875	196.037	166.709
48	071123_0_R00X152Y088	12.8755	27.3605	25.7511	42.382	39.6996	57.9399	56.867	89.0558	148.605	133.584
49	071123_0_R00X152Y090	21.2811	32.664	27.7149	93.5378	146.493	79.1854	60.3789	99.4767	163.815	174.208
50	071123_0_R00X152Y092	21.0364	22.1735	22.742	105.75	117.121	31.8388	57.4236	69.3632	206.952	237.086
51	071123_0_R00X154Y090	19.1879	18.3157	45.3532	70.6463	99.4281	75.8793	72.3906	80.2402	122.105	134.315
52	071123_0_R00X154Y092	10.5013	25.8494	23.4261	50.0833	37.1586	40.3898	47.6599	54.9301	79.9717	114.707
53	071123_0_R00X156Y090	23.6236	21.5143	36.7009	76.3548	115.587	47.2472	57.3715	89.4321	193.629	218.94
54	071123_0_R00X156Y092	13.5108	10.9126	18.7073	109.646	49.3665	72.231	94.0561	71.7113	144.462	183.955
55	071123_0_R00X156Y094	22.2561	19.4741	28.7474	52.8582	107.571	31.5294	124.263	115.917	236.471	204.014
56	071123_0_R00X156Y096	25.5867	33.4596	14.7616	81.6807	101.363	81.6807	115.14	66.9191	84.633	106.283
57	071123_0_R00X158Y092	15.8937	16.4613	24.9757	44.2752	22.1376	47.1133	64.1423	97.6325	186.75	221.944
58	071123_0_R00X158Y094	7.40787	35.9811	17.9905	67.7291	135.458	66.6708	58.2047	180.964	155.565	139.691
59	071123_0_R00X158Y096	9.55467	22.0492	21.6817	102.896	49.9783	66.8827	92.9743	113.921	153.61	207.63
60	071123_0_R00X158Y098	23.4534	48.7831	28.1441	77.8853	101.319	51.5975	60.9789	58.1645	121.958	174.493
61	071123_0_R00X158Y100	21.5343	3.58905	19.1416	45.4614	82.5483	25.1234	65.7993	88.53	203.38	197.398
62	071123_0_R00X158Y102	23.4076	12.8742	31.6003	98.312	80.7563	80.7563	80.7563	102.994	122.89	152.15
63	071123_0_R00X158Y104	14.6501	19.8208	13.7884	102.551	91.3479	38.7798	130.128	103.413	175.802	162.875
64	071123_0_R00X160Y092	8.0751	19.9186	24.7636	99.0546	39.2988	36.6071	97.4395	80.751	145.89	123.818
65	071123_0_R00X160Y094	15.9116	26.2887	30.4396	51.8857	80.2498	87.1679	62.9546	65.7218	214.461	150.814
66	071123_0_R00X160Y096	8.60331	24.7345	54.8461	88.1839	74.2035	61.2986	74.2035	88.1839	176.368	153.784
67	071123_0_R00X160Y098	17.4303	24.4024	43.9243	85.757	82.271	59.9602	64.8407	80.8766	138.745	119.223
68	071123_0_R00X160Y100	11.8858	27.5921	38.0535	108.671	52.6373	78.1069	78.5314	156.638	115.887	128.622
69	071123_0_R00X160Y102	15.9645	29.6484	19.3855	37.6307	90.0855	86.6646	95.7871	115.173	117.453	178.75
70	071123_0_R00X160Y104	7.84482	16.25	23.5345	37.5431	59.3965	27.4569	45.9482	52.6724	75.6465	85.7327
71	071123_0_R00X160Y106	14.4514	18.8497	31.4162	83.567	42.0977	70.3722	71.0006	59.6907	106.815	140.116
72	071123_0_R00X162Y092	11.268	45.877	41.8527	145.68	47.4867	90.9482	111.071	100.607	174.655	143.265
73	071123_0_R00X162Y094	23.7069	25.6824	27.858	84.9496	119.851	65.1939	55.9745	38.2188	69.145	67.8279
74	071123_0_R00X162Y096	14.8063	14.8063	23.6901	75.5123	44.419	53.3028	69.5898	105.125	156.947	198.405
75	071123_0_R00X162Y098	8.33317	17.3074	29.4866	93.5879	44.8709	46.794	98.075	70.5114	149.358	190.381

Fig. 10.7 Example representation of a peak list generated by Spec Align and Microsoft Excel software. Each column represents a spectrum and each line represents peak intensity [7]

	A	B	C	D	E	F	G
1							
2	**Component Score**						
3		PC1	PC2	PC3	PC4	PC5	···
4	Spectrum (X058Y104)	-10.13370794	-0.445522876	-2.694187973	3.505954923	-1.365223441	···
5	Spectrum (X058Y105)	-9.229694994	-0.567799789	0.237015377	1.116560832	-1.233019483	···
6	Spectrum (X058Y106)	-5.602620225	-0.656359247	1.143176137	-0.865466241	-0.732606489	···
7	Spectrum (X058Y107)	-6.270811331	-2.03169119	-0.285734608	-1.819995303	-0.650038451	···
8	Spectrum (X058Y108)	-7.66653915	0.364231772	-0.655977851	3.500619798	-0.957857607	···
9	Spectrum (X058Y109)	-5.794829016	-0.370002515	1.224927625	2.69823407	-0.576784283	···
10	Spectrum (X058Y110)	-6.095335164	-0.961580719	1.025728904	2.007179238	-0.679957134	···
11	Spectrum (X058Y111)	-10.74294884	-0.105700353	-1.750048371	3.79086667	-1.19753972	···
12	Spectrum (X058Y112)	-7.764753996	-0.612978725	1.471680849	0.700552849	-0.801859914	···
13	Spectrum (X058Y113)	8.627123127	-4.905947292	-0.942587556	-1.087932962	-6.388071894	···
14	Spectrum (X058Y114)	:	:	:	:	:	

Fig. 10.8 Example representation of a PCA result performed by Excel add-ins. Component scores are calculated for each mass spectrum, according to each principal component

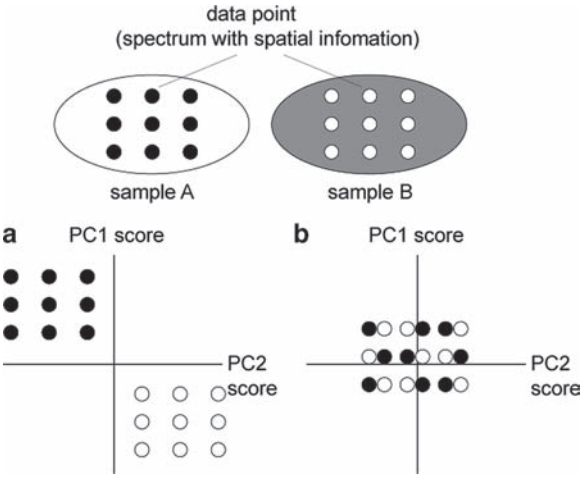


Fig. 10.9 Example of data interpretation of IMS-linked PCA. In this study, dots seen in the 2D plot represent the *case*, i.e., spectrum from distinct data points. If dots from distinct sample are separated (**a**), it means that the molecular expression patterns of these two regions were statistically distinct from each other. If not, PCA failed to extract the statistical differences between the two populations (**b**)

Note: We additionally noted that other artificial factors, such as variations in matrix application procedures, could be reflected as the main difference between the two groups in PC1. To avoid this, experiments should be performed in as identical conditions as possible (vis-à-vis matrix application, IMS measurement, etc.).

10.4.3.2 Interpretation of Factor Loading

An analysis of factor loading plot would identify peaks that were differentially expressed between two samples. A component score defined for each spectrum is a sum of the factor loading value, multiplied by each peak intensity. Therefore, when

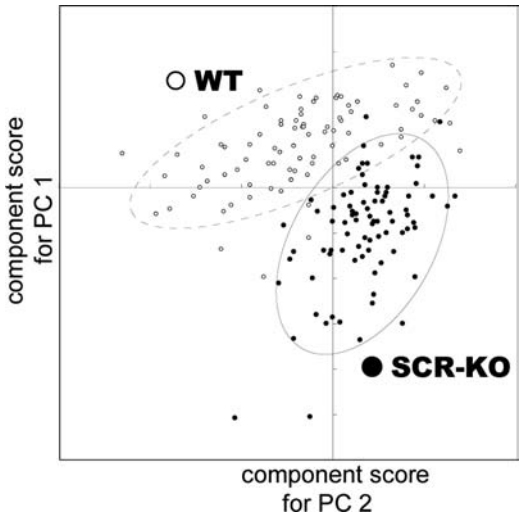


Fig. 10.10 Component scores for PC1 and PC2 are plotted on the *x*-axis and *y*-axis, respectively. Each spot represents a mass spectrum. *Open* and *filled spots* indicate spectra obtained from *WT* and *SCR-KO* mice, respectively

	A	B	C	D	E	F	G
1	**Factor Loading**						
2		PC1	PC2	PC3	PC4	PC5	···
3	m/z 1119	0.071506233	-0.005891377	-0.326683397	0.292888681	0.198868401	···
4	m/z 2153	-0.1486389	0.223434933	-0.489005528	0.230359915	0.091340758	···
5	m/z 9756	0.182920729	-0.152870098	-0.451659506	0.269897412	0.144066523	···
6	m/z 9809	-0.333347528	0.096143037	-0.345039238	0.089061045	0.062427567	···
7	m/z 10082	-0.294288151	-0.002173624	-0.418962896	-0.086867098	-0.038617682	···
8	:	:	:	:	:	:	
9	eigenvalue	17.89857465	4.949642519	4.509393439	2.764140748	2.075291342	
10	proportion	22.09700574	6.110669777	5.567152394	3.412519442	2.562088077	
11	cumulative proportion	22.09700574	28.20767551	33.77482791	37.18734735	39.74943543	

Fig. 10.11 Factor loading value is calculated for each mass peak

numbers (=*m*) of mass peaks were used in the analysis, the component score will be as follows:

$$\mathrm{ScorePC1}(x, y) = \sum_{n=1}^{m} \mathrm{load}(n) \times \mathrm{Int.}(n),$$

where ScorePC1(*x*, *y*) is the component score against PC1, obtained from (*x*, *y*); load(*n*) is the factor loading value against a mass peak for *n*; Int.(*n*) is the mass peak intensity for *n*; and *m* is the number of mass peaks used for calculation.

According to this equation, in the spectra from Scr-KO mice, the mass peak with large negative value regarding PC2 factor loading was supposed to be intense. On the other hand, in the WT sample, it was supposed that such peak intensities would be small. In other words, such mass peaks are suggested to differentiate samples derived from WT and Scr-KO tissue.

The factor loading value is calculated for each mass peak. In Fig. 10.11, the factor loading values for PC1 and PC2 are plotted on the *x*-axis and *y*-axis, respectively. Each spot indicates a distinct mass peak. Such a graph makes it very easy to find the

peaks with the intended factor loading value against each PC. Because peaks that have negative loading values regarding PC2 are supposed to show different distributions between WT and Scr-KO, we picked up a mass peak at *m/z* 7,420 and obtained a distribution image (Fig. 10.12).

As a result, we found that this ion was highly expressed in the striatum of WT mice whereas it was expressed at remarkably low levels in the Scr-KO striatum (Fig. 10.13, arrowheads). Intriguingly, in olfactory bulb, there is no significant difference in the expression levels of *m/z* 7,420, between the two samples (Fig. 10.13, arrows). Furthermore,

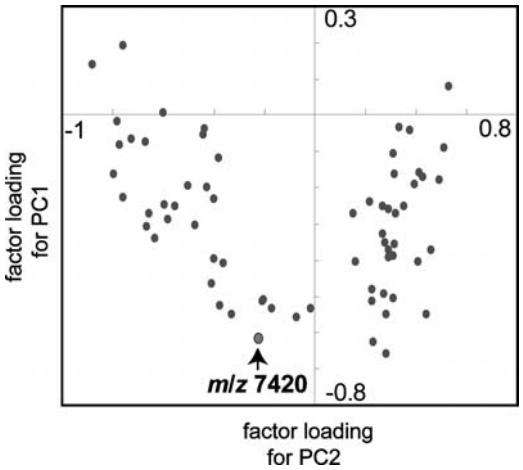


Fig. 10.12 Factor loading values for PC1 and PC2 are plotted on *x*-axis and *y*-axis, respectively. Each spot indicates the distinct mass peak. The peak (*m/z* 7,420) was chosen because it has large negative loading values regarding PC2

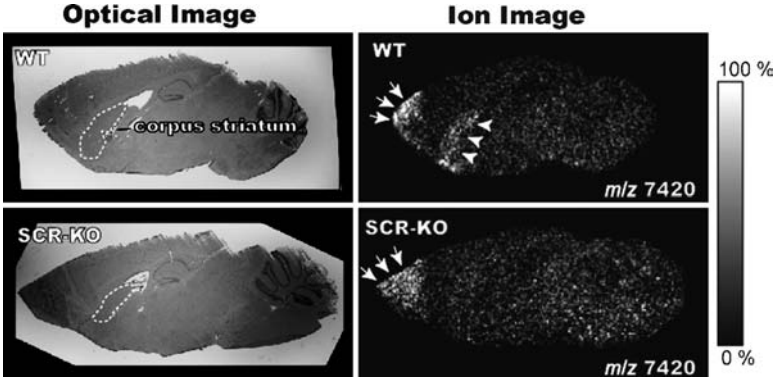


Fig. 10.13 H&E-stained brain sections prepared from WT and SCR-KO mice, with distribution of *m/z* 7,420. It was highly expressed in the striatum of WT mice but was expressed at remarkably low levels in Scr-KO striatum tissue (*arrowheads*). On the other hand, there was no significant difference in the expression levels of *m/z* 7,420 in the olfactory bulb, between the two samples (*arrows*) (Reprinted from Yao et al., Proteomics 8:3692–3701.)

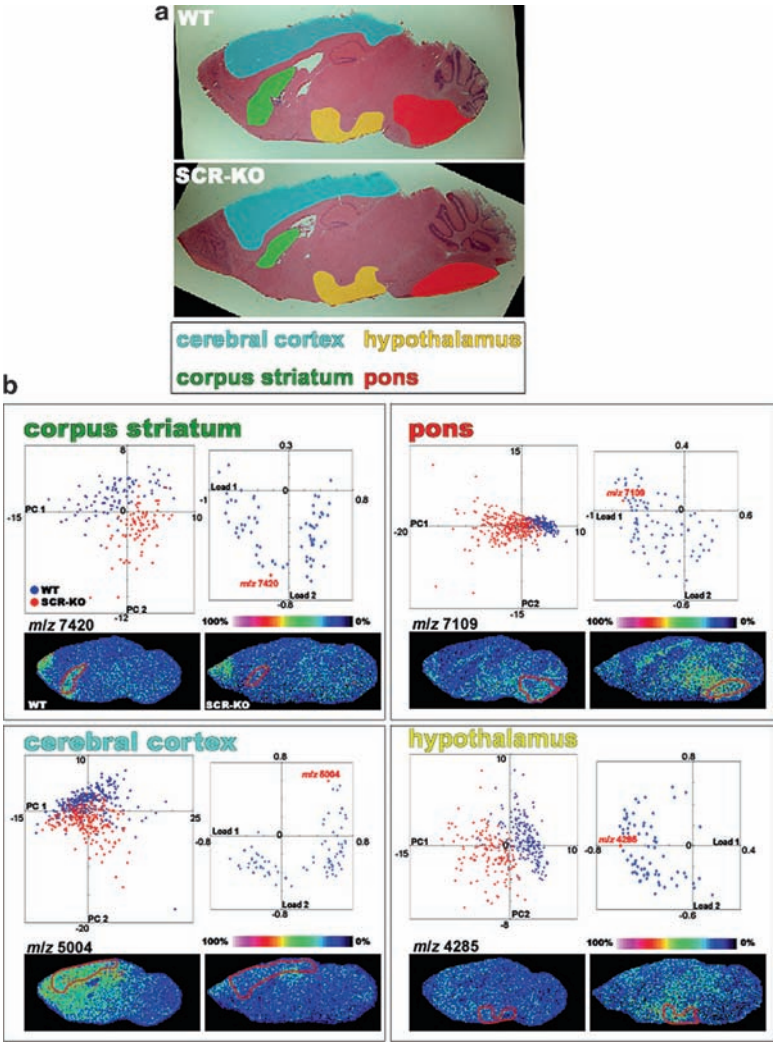


Fig. 10.14 PCA-linked IMS revealed abnormal expressions/suppressions of proteins in various regions of SCR-KO mouse brains. **a** H&E-stained images of WT and SCR-KO mouse brains. The regions focused upon during MS imaging analysis are indicated in *colors* as keyed below the pictures in (**a**). **b** Distributions of principal component scores of mass spectra, from various brain regions (*left spray graphs:* WT, *blue*; KO, *red*) and the factor loading plot (*right graphs*). The signal intensities of the mass spectra of the substances, with indicated *m/z*, are shown in the reconstructed images of the mouse brains analyzed with MS imaging [7] (Reprinted from Yao et al., Proteomics 8:3692–3701.)

through the same procedure, we were able to successfully identify the altered protein expression patterns in various selected brain regions (Fig. 10.14).

Note: We finally note that it is important to pay attention to the contribution of each PC to determine how many PCs are to be used in analysis. Proportion values represent

the degree of a particular PC's contribution to the entire dataset (the values were seen in the 10th column of Fig. 10-10). For example, the proportion value of PC1 in this analysis is 22%, indicating that PC1 reflects 22% of the entire dataset. Any PC with too low a proportion value should be excluded from the analysis because we know which PC reflects data significantly in this manner. It is also important to note that PC1 often reflects variations in tissue sample preparation and matrix coating.

10.5 Conclusion

The volumes of IMS datasets continue to increase because of current improvements to IMS with regard to high resolution [20], three-dimensional (3D) imaging [21], and reconstruction from 3D mass spectra containing ion drift times in ion mobility MS [22]. Data analysis of such large datasets will increasingly depend on statistical analysis and will presumably be done automatically by software programs.

In this chapter, we showed a simple protocol that is needed to identify the differences between two samples. By using this minimal procedure, we were able to identify a molecule which differed between WT and Scr-KO tissue samples. Again, we would recommend that IMS experimenters perform statistical analysis by themselves, because undertaking such exercises will provide useful experience in understanding more complicated analyses, especially among biologists and clinicians.

References

1. Altelaar AF, Luxembourg SL, McDonnell LA, et al. (2007) Imaging mass spectrometry at cellular length scales. Nat Protocols 2:1185–1196
2. McCombie G, Staab D, Stoeckli M, et al. (2005) Spatial and spectral correlations in MALDI mass spectrometry images by clustering and multivariate analysis. Anal Chem 77:6118–6124
3. Plas RV, Moor BD, Waelkens E (2007) Imaging mass spectrometry-based exploration of biochemical tissue composition using peak intensity weighted PCA. In: 2007 IEEE/NIH Life Science Systems and Applications Workshop, pp 209–212
4. Yanagisawa K, Shyr, Y, Xu, B J, et al. (2003) Proteomic patterns of tumour subsets in non-small-cell lung cancer. Lancet 362:433–439
5. Mantini D, Petrucci F, Del Boccio P, et al. (2008) Independent component analysis for the extraction of reliable protein signal profiles from MALDI-TOF mass spectra. Bioinformatics 24:63–70
6. Prideaux B, Atkinson SJ, Carolan VA, et al. (2007) Sample preparation and data interpretation procedures for the examination of xenobiotic compounds in skin by indirect imaging MALDI-MS. Int J Mass Spectrom 260:243–251
7. Yao I, Sugiura Y, Matsumoto M, et al. (2008) In situ proteomics with imaging mass spectrometry and principal component analysis in the Scrapper-knockout mouse brain. Proteomics 8:3692–3701
8. Hanselmann M, Kirchner M, Renard BY, et al. (2008) Concise representation of mass spectrometry images by probabilistic latent semantic analysis. Anal Chem 80(24):9649–9658

9. Fowlkes EB, Mallows CL (1983) A method for comparing two hierarchical clusterings. J Am Stat Assoc 78:553–584
10. Deininger SO, Schürenberg M, Suckau D, et al. (2007) Class imaging: classification of breast cancer sections by MALDI tissue imaging. Poster presentation in HUPO
11. Denkert C, Budczies J, Kind T, et al. (2006) Mass spectrometry-based metabolic profiling reveals different metabolite patterns in invasive ovarian carcinomas and ovarian borderline tumors. Cancer Res 66:10795–10804
12. Lapolla A, Ragazzi E, Andretta B, et al. (2007) Multivariate analysis of matrix-assisted laser desorption/ionization mass spectrometric data related to glycoxidation products of human globins in nephropathic patients. J Am Soc Mass Spectrom 18:1018–1023
13. Wong WHJ, Cagney G, Cartwright HM (2005) SpecAlign: processing and alignment of mass spectra datasets. Bioinformatics 21:2088–2090
14. Yao I, Takagi H, Ageta H, et al. (2007) SCRAPPER-dependent ubiquitination of active zone protein RIM1 regulates synaptic vesicle release. Cell 130:943–957
15. He L, Lu XY, Jolly AF, et al. (2003) Spongiform degeneration in mahoganoid mutant mice. Science 299:710–712
16. Sugiura Y, Shimma S, Setou M (2006) Thin sectioning improves the peak intensity and signal-to-noise ratio in direct tissue mass spectrometry. J Mass Spectrom Soc Jpn 54:4
17. Schwartz SA, Reyzer ML, Caprioli RM (2003) Direct tissue analysis using matrix-assisted laser desorption/ionization mass spectrometry: practical aspects of sample preparation. J Mass Spectrom 38:699–708
18. Norris JL, Cornett DS, Mobley JA, et al. (2006) Processing MALDI mass spectra to improve mass spectral direct tissue analysis. Int J Mass Spectrom 260:212–221
19. Sugiura Y, Konishi Y, Zaima N, et al. (2009) Visualization of the cell-selective distribution of PUFA-containing phosphatidylcholines in mouse brain by imaging mass spectrometry. J Lipid Res (in press)
20. McDonnell LA, Piersma SR, Maarten Altelaar AF, et al. (2005) Subcellular imaging mass spectrometry of brain tissue. J Mass Spectrom 40:160–168
21. Andersson M, Groseclose MR, Deutch AY, et al. (2008) Imaging mass spectrometry of proteins and peptides: 3D volume reconstruction. Nat Methods 5:101–108
22. McLean JA, Ridenour WB, Caprioli RM (2007) Profiling and imaging of tissues by imaging ion mobility-mass spectrometry. J Mass Spectrom 42:1099–1105