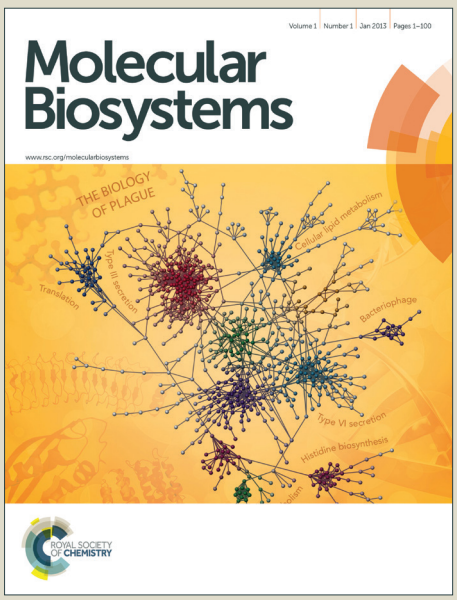
Molecular BioSystems

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems

Molecular BioSystems

ARTICLE

A new approach combining LC-MS and multivariate statistical analysis for revealing changes in histone modification levels.

Raphaël Bilgraer[a], Sylvie Gillet[a*], Sophie Gil[b], Danièle Evain-Brion[b], Olivier Laprévote[a,c]

Cite this: DOI: 10.1039/x0xx00000x

Received 00th January 2012,
Accepted 00th January 2012

DOI: 10.1039/x0xx00000x

www.rsc.org/

While acting upon chromatin compaction, histone post-translational modifications (PTMs) are involved in modulating gene expression through histone-DNA affinity and protein-protein interactions. These dynamic and environment-sensitive modifications are constitutive of the histone code that reflects the transient transcriptional state of the chromatin. Here we describe a global screening approach for revealing epigenetic disruption at the histone level. This original approach enables fast and reliable relative abundance comparison of histone PTMs and variants in human cells within a single LC-MS experiment. As a proof of concept, we exposed BeWo human choriocarcinoma cells to sodium butyrate (SB), a universal histone deacetylase (HDAC) inhibitor. Histone acid-extracts (n=45) equally representing 3 distinct classes, Control, 1 mM and 2.5 mM SB were analyzed using ultra-performance liquid chromatography coupled with a hybrid quadrupole time-of-flight mass spectrometer (UPLC-QTOF-MS). Multivariate statistics allowed us to discriminate control from treated samples based on differences in their mass spectral profiles. Several acetylated and methylated forms of core histones emerged as markers of sodium butyrate treatment. Indeed, this untargeted histonomic approach could be a useful exploratory tool in many cases of xenobiotic exposure when histone code disruption is suspected.

Introduction

Histones are some of the most heavily modified proteins in the cell nucleus, and form the nucleosomes together with DNA.[1] A wide range of post-translational modifications (PTMs) has been identified so far, the most studied still being acetylation, mono-, di- or trimethylation and phosphorylation. These modifications help chromatin dynamic changes, thus gene expression[2–4] and some hallmarks are generally associated with heterochromatin while others seem to be more specific of euchromatin.[5,6] Revealing PTM patterns is a tricky task as these various modifications act in a combinatorial mode, commonly called the histone code.[7] Besides, histone covalent modifications act also via protein-protein interactions and thus take part of some specific protein recruitment.[8] It is also well-known that histone modification levels vary both locally and globally throughout various events like cell cycle, development,[9] pathogenesis or in response to exogenous stimuli.[10] Thereby, disruption of the histone code can have serious consequences on gene expression and cellular homeostasis. As a result, numerous studies

have dealt with histone PTMs characterization. Traditionally, chromatin studies have rather involved immunoassay methodologies like western blotting, immunoprecipitation or immunofluorescence. However, these approaches share the major disadvantage of being targeted, as they all closely rely on PTM-specific antibodies. This kind of technologies points out only a few specific modifications and cannot provide a global combinatorial PTMs image. Additionally, antibodies efficiency might be affected by multiple PTMs,[11] and histone characterization can become really tough due to the existence of sequence variants, particularly in the case of H2A and H2B subtypes. Analytical approaches are nowadays more and more employed, and mass spectrometry becomes a key tool for histone PTMs characterization.[12] The vast majority of studies has been mainly done at the peptide level by the so-called bottom-up mass spectrometry approach.[13] Nevertheless, as immunoassay methods, this kind of strategy prevents from any combinatorial aspect of the histone PTMs and thus can't help in fully understanding biological function. Part of the peptidic information can indeed be lost in the case of trypsinolysis as histones are small highly basic proteins. In such context, some lysine derivatisation methods have been proposed to improve the peptide recovery rate.[14,15] Over the last decade, the top-down approach has tremendously emerged in the histone field in order to overcome the lack of combinatorial information observed with the other methods and many studies are presently carried out at the intact protein level.[16–18] Most of these studies rely on liquid chromatography combined with Fourier transform-mass spectrometry and electron transfer dissociation (ETD) or electron capture dissociation (ECD) for ensuring accurate variant assignment as well as PTMs localisation.[19,20] Some other studies, namely the middle-down ones, involve very large peptides rather than short tryptic peptides or intact proteins.[21] Some have even pioneered the use of ultra-high performance liquid chromatography to limit the required amount of sample and improve histone separation.[22] These strategies have the advantage of preserving the combinatorial information and help in biological features interpretation. More recently, some studies used mass spectrometry in combination with *in vivo* labeling in order to get access to quantitative data on histone PTMs dynamic.[23,24] Lysine acetylation of histones is due to lysine/histone acetyl transferases (KATs/HATs), and is commonly associated with a loosening of DNA-histone interaction, thus increasing the gene transcription rate. Conversely, the interaction tightness can be restored by means of histone deacetylases (HDACs), leading to a repressive state of the chromatin. Hypo- and hyper-acetylation have been correlated to either severity or good prognosis in various cancers and non-cancer human diseases in a tissue-specific manner,[25–27] leading researchers to investigate whether the HATs/HDACs balance disruption could be exploited as a new cancer therapy approach. By inhibiting deacetylation of histone tails, HDAC inhibitors (HDACIs) induce chromatin remodeling and modulate the expression of target genes related to malignant phenotypes.[28] Moreover, the resulting histone hyperacetylation can affect the activity of several non-histone proteins and regulate different molecular processes (mitosis, DNA replication, DNA repair).[29] For such reasons, HDACIs, mostly in combination with other drugs, are considered as promising candidates for the treatment of cancer and other various human diseases. Recent studies relying on mass spectrometry explored the histone acetylation from a quantitative and a dynamic point of view at the peptide level.[30,31] In order to remove the difficulties associated with the digestion step and with the peptide characterization and quantitation, we put our efforts on establishing a new analytical approach for the differential analysis of histone PTMs at the intact protein level by combining the use of UPLC-QTOF-MS and multivariate statistical analysis (MVA). Supervised and unsupervised statistical methods have been widely used in different engineering or scientific disciplines for several years, and more recently in -omic sciences.[32] These advanced statistical techniques belong to the pattern recognition field which aims to teach machines how to observe and distinguish patterns in the data and make the decision to classify samples (*i.e.* spectra) into different categories.[33]

Here, we focused our attention on the differential analysis of acetylated and methylated histones in the BeWo choriocarcinoma cell line either untreated or treated with 1 or 2.5 mM sodium butyrate (SB), a well-known HDACI. We applied MVA (PCA, PLS-DA, OPLS-DA) to the experimental LC-MS data in order to extract the most informative features within high dimensional data sets.[34] Our global strategy allowed us assessing the global acetylation level of histones and we have been able to confirm the SB dose-dependent acetylation level on a reliable statistical basis. This original approach might be rather useful as first-line investigation method for studying histone PTMs levels, one of the main advantages being the straightforwardness of the experimental process from histone extract to acetylation level evaluation without any protein or peptide sequencing.

Results and discussion

UPLC-MS profiles of intact proteins

Since histones may be present in small amounts depending on the number and type of cells from which they are extracted, we've put our efforts on working with histone amounts as little as 1.5 µg, while other methods dealt with tens of micrograms. Each proteomic profile consisted of 10,004 distinct *m/z* values. After analysis of both chromatograms (Fig. 1) and XICs, we noticed that core histones eluted between 10 and 16 minutes. We thus removed from the *X*-matrix variables whose retention times were out of this range, leading to a refined dataset containing 8,537 *m/z* values. We estimated the global level

of precision of our method to 18.8% by calculating the median %RSD for the QC replicates. This result was in agreement with the FDA guidelines for biomarker studies which recommend an overall analytical precision under 20%.[35] We decided to use centroid data for the multivariate statistical analysis since deconvoluting intact protein spectra may lead to some artifacts that would have distorted the chemometric analysis.

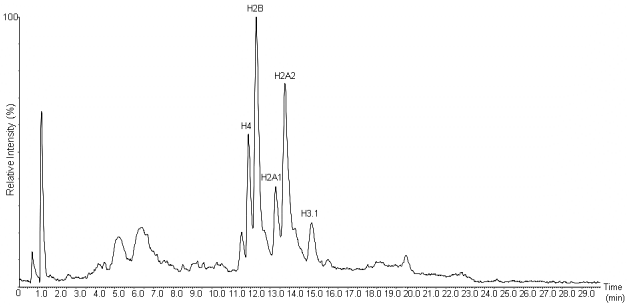


Fig. 1: Total ion chromatogram (TIC) of unfractionated acid-extracted histones. About 1.5 µg of total proteins were injected.

Data normalisation

Pattern recognition techniques are used to explore structure in datasets by modeling sources of variance between samples. Hence, it is necessary to correct for sources of systematic variation between spectral profiles before comparing them.[36,37] Two data normalisation steps were sequentially applied to improve the performance of MVA.[38] First, a global normalisation was performed to remove the unwanted variability that affects all features of a spectrum. It consisted in calculating the median of variables for each spectrum as a rescaling coefficient.[39] A $\log_{10}$ transformation was applied to make skewed distribution of the data more symmetric. Next, the data were Pareto-scaled to adjust variables magnitude and reduce their range of variance across samples.[40] The Pareto scaling is known to give better results than other scaling methods (*e.g.* unit-variance) on MS data since it better preserves the original structure of the data.[41]
Boxplots in Fig. 2 summarize the effect of normalisation on the distribution of variable intensities. Kernel density estimation of the data before and after normalisation is also represented (Fig. 3), attesting for the normal distribution of the data.

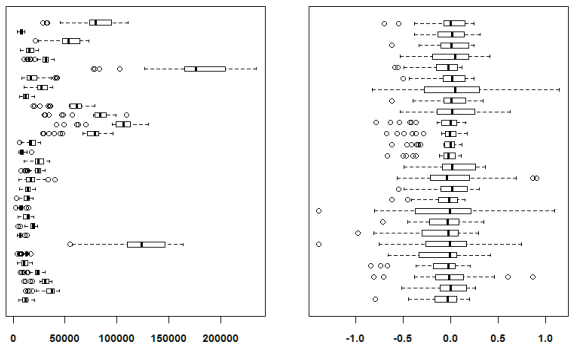


Fig. 2: Boxplots for 25 randomly selected features out of the 8,537 before (left) and after (right) data normalisation. Variable intensity is represented along the x-axis.

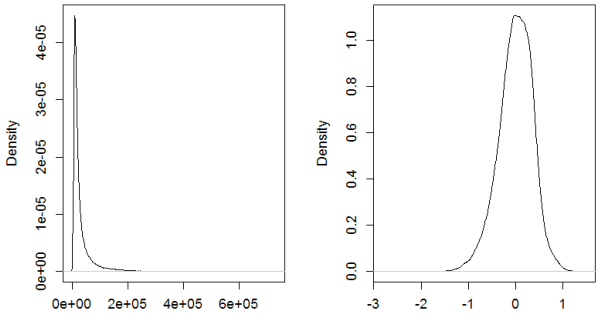


Fig. 3: Kernel density plots for the 8,537 features across all samples before (left) and after (right) data normalisation. Variable intensity is represented along the x-axis.

Unsupervised statistical analysis

Hierarchical cluster analysis. Once data have been properly preprocessed, hierarchical clustering was first applied to the training set to agglomerate samples with similar profiles of variables intensity.[42] The heatmap in Fig. 4 represents the top 250 variables ranked by analysis of variance (ANOVA). The results show that all samples belonging to the same class were correctly grouped together. The treelike structure of the dendrogram highlighted two main clusters of variable intensities represented by its two longest branches (maximum dissimilarity according to Euclidean distance). According to the color gradient, the intensity differences between groups are substantial. The two main branches of the dendrogram represent the variables that were up- (left branch) or down- (right branch) regulated by the SB treatment. This first analysis allowed us to easily detect natural clusters in the data, although it didn't facilitate extraction of discriminant variables among the 8,537 features.

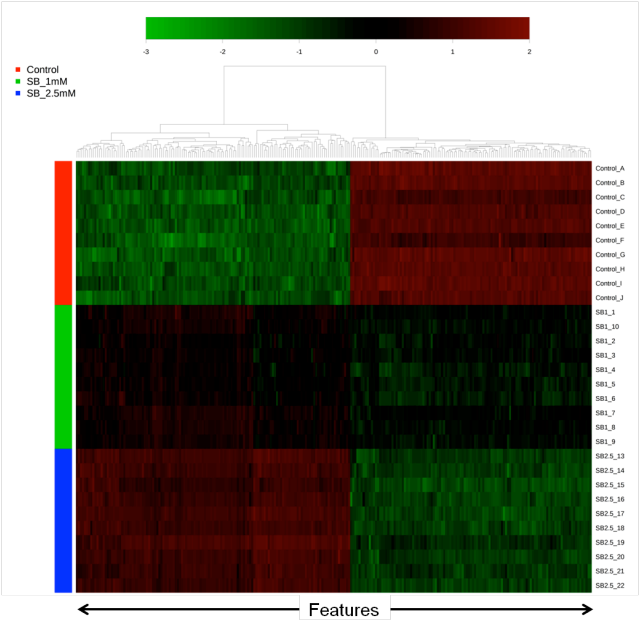


Fig. 4: Hierarchical cluster analysis and heatmap visualisation of top 250 variables (x-axis) ranked by ANOVA. The biological replicates (*n*=30) for each sample class within the training set are represented along the y-axis. The color code was used to represent log-scaled intensities of features between -2 (*green*) and +2 (*red*), showing the up- or downregulation of variables according to the SB dose.

PCA overview. To further explore natural separation between histone profiles and reduce data dimensionality, the training set composed of 10 samples per condition was submitted to principal component analysis (PCA). The number of significant components was estimated using internal cross-validation with seven exclusion groups,[43] giving a three-component PCA model accounting for 35,5 % of the total variance (R^2X, cumulative variance of all the X-matrix explained by all extracted components). The resulting scores plot was used to identify trends, groups and potential outliers within the data. Fig. 5 shows that PCA analysis succeeded in clustering all samples according to the SB dose. The clear separation between control, 1 mM and 2.5 mM SB samples indicated that the major source of variance was unambiguously related to the treatment conditions. This confirms that sodium butyrate induced substantial changes in histone PTMs patterns of exposed samples. Moreover, a good discrimination between 1 mM and 2.5 mM SB groups was observed, connoting a dose-response relationship. The PCA scores plot also demonstrated the tight clustering of QC samples, meaning that analytical variability was acceptable. Neither severe nor moderate outlier was detected using the *Distance to the Model in X space* (DModX).[44]

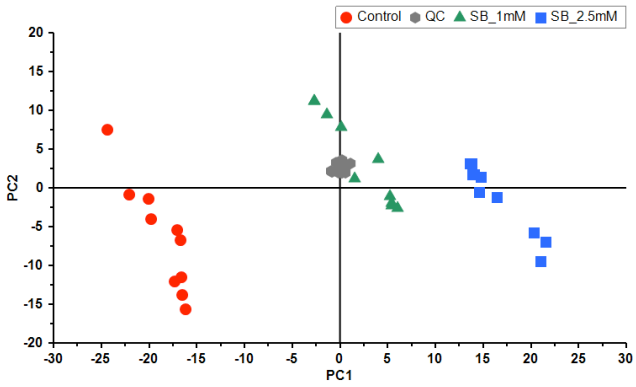


Fig. 5: Principal component analysis scores plot of the normalised training set. The four classes are represented by different symbols (O = Control, Δ = SB 1 mM, □ = SB 2.5 mM and ⬢ = QC). Samples are well separated on PC1 according to their class membership, and the QC samples are tightly clustered in the middle, accounting for the good quality of the data.

Supervised analysis

PLS-DA modelling of the three classes. In our classification problem, supervised methods are more suitable to accurately model the relationship between SB treatment and spectral profiles. PLS-DA classification was first applied to the training set previously used for PCA analysis. Samples were labeled in belonging to one of the 3 classes (1 = Control, 2 = 1 mM SB, 3 = 2.5 mM SB). A model was considered predictive if the Q^2 (cross-validated R^2, measure of the predictive power) regression line intercept resulting from the permutation test was negative.[45] The number of significant components was determined as previously described for PCA. The final three-component model had an R^2 of 0.99 and a Q^2 of 0.94. The proximity of R^2 and Q^2 values ensures that the model was not overfitted. The scores plot (Fig. S1a) revealed that each class was well separated, suggesting that the PLS-DA model successfully discriminated samples according to their treatment. This model was internally validated both by CV-ANOVA (*p*-value = $1.8\ 10^{-14}$) and by the permutation test (999 permutations gave a negative Q^2 intercept).[46] Nevertheless, the best way for accurately estimating the predictive power of PLS-DA models is to test their ability to classify unknown samples coming from a blind test set. Then a test set consisting of 15 samples (5 per class) was projected into the PLS-DA space previously defined by the training set (Fig. S1b). Despite an adequate distribution of all the samples between the 3 classes, binary classification data (Table S1) indicated that within the 2.5 mM SB group, 3 samples were predicted as belonging to the 1 mM SB group. This resulted in 80% of accurate predictions. Although this model succeeded in discriminating between control and treated samples, it has not been able to accurately predict the SB dose. The large

number of shared variables between the two SB-treated groups could explain this slight lack of prediction power. Therefore, we focused our attention on two-class comparisons using OPLS-DA models to increase classification accuracy and facilitate models interpretation.

Two-class OPLS-DA models. The first OPLS-DA model was built using a training set including 10 control and 10 samples treated with 1 mM SB. The model had one predictive and one orthogonal component (1+1), and its descriptive statistics were as follows: $R^2 = 0.99$, $Q^2 = 0.97$ and CV-ANOVA p-value = $4.5\ 10^{-11}$. The corresponding scores plot shown in Fig. 6a exhibited a clear separation between the two classes on the predictive component. This was confirmed by plotting the observed versus predicted group values (Fig. S2), where all samples were correctly spotted on either side of 0.5 on the x-axis.[47] The significant OPLS-DA model was then applied to the test set containing 10 samples. The resulting projection is shown in Fig. 6b. All samples were correctly classified, giving 100 % of accurate predictions (Table S2). The OPLS-DA model for the 2.5 mM SB dose was build according to the same procedure, leading to a 1+1 component model with $R^2 = 0.99$, $Q^2 = 0.98$ and CV-ANOVA p-value = $9.3\ 10^{-13}$ (Fig. S3, S4, S5). As done previously, the statistically significant model was validated by an external test set, resulting in 100 % of accurate predictions (Table S3).

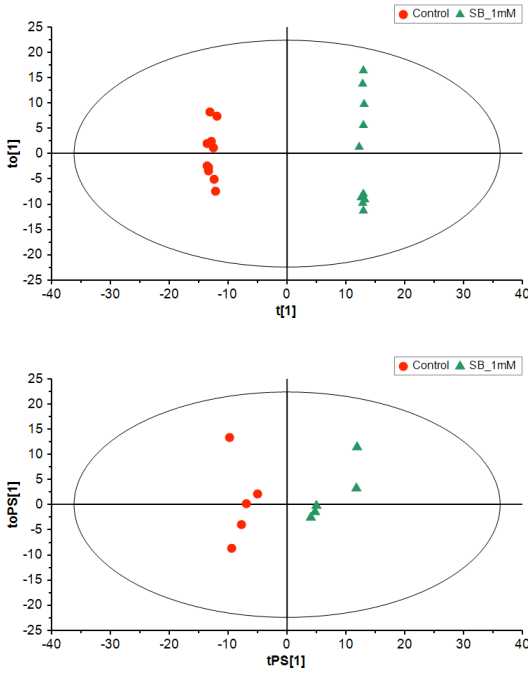


Fig. 6: Classification of control and 1 mM SB samples: OPLS-DA scores plot of the training set (up, *n*=20) and projection of the test set (down, *n*=10).

Discriminant histone forms associated with SB treatment

Selection of discriminant variables was achieved using the VIP scores procedure for each validated OPLS-DA model. Based on a 1.5 cutoff value, 155 features out of the 8,537 were selected for the 1 mM SB model and 115 for the 2.5 mM SB model. Moreover, all the absolute correlation coefficients (p(corr)) for these extracted variables were over 0.6, where a cutoff of 0.4-0.5 is often applied.[48] These high numbers of discriminant variables were partially due to redundancy related to the distribution of the multiply charged ions detected for each protein (Fig. 7).

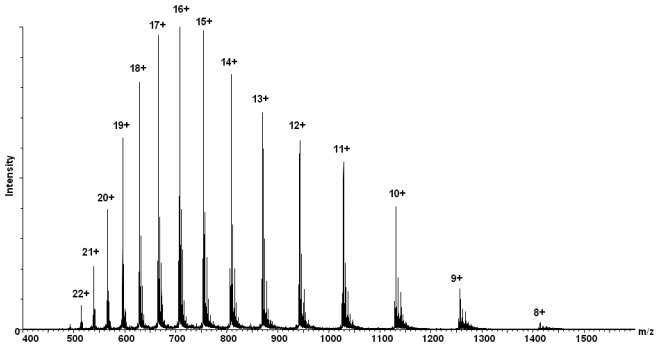


Fig. 7: Continuum mass spectrum of intact histone H4 showing the distribution of the charge states from 8+ to 22+.

We then refined the two lists of variables by retaining only the most intense charge state for each discriminant protein form. Continuum mass spectra were deconvoluted using the maximum entropy algorithm (MaxEnt I)[49] leading to simplified spectra of intact proteins. These spectra were manually analysed to assign the molecular masse (MM) of each selected variable with an average 0.5 Da error in mass (approximately 45 ppm). Most probable histone forms were further identified by matching experimental average masses with theoretical ones with the help of the free TagIdent tool[50] available on the ExPaSy server. This software provided us a list of proteins that met the selected parameters (taxonomy: Homo Sapiens, pI range = 10.0 - 12.0 and MM = observed average mass ± 0.005%). Considering these parameters, only histones were returned by TagIdent, and a maximum of two different histone forms were proposed at a time. Thus, identifications were proposed based on both TagIdent results and lists of referenced PTMs on HIstome[51] and UniprotKB/Swiss-Prot[52] on-line databases. Nevertheless, one has to bear in mind that working at the intact protein level can lead to some identification ambiguities that would sometimes require protein sequence determination or ultra-high resolution capabilities. Regarding the quantitative aspect, previous studies have shown that working at the protein or the peptide level led to similar results.[22,53]

Due to the quite high complexity of H2A and H2B variant patterns, further complicated by the combinatorial aspect of

their PTMs, we have not always been able to assign intact masses to specific proteoforms with confidence.[54] We thus decided to mainly focus our detailed comments on H3 and H4 histones. We deliberately associated a mass increment of +42 Da with lysine acetylation since variables found discriminant were resulting from a HDACI treatment. Similarly, we associated a mass increment of +28 Da with either two lysine monomethylation or lysine dimethylation rather than the low-abundant lysine formylation.[55] The resolution of our Q-TOF instrument is anyway far below the 300,000 required to distinguish these particular isobaric modifications at the intact protein level.

At 1 mM SB, up to 17 histone forms were identified as discriminant between 1 mM SB-treated and control samples. These discriminant variables are summarized in table 3. Within them, 10 variables corresponded to H4 modified forms, 6 putatively corresponded to H2B1 and only one corresponded to H2A1. As shown in Table 1, the most heavily acetylated forms were upregulated by the 1 mM SB dose, while the non- or weakly acetylated ones were downregulated by the SB treatment. Since the global amount of histones incorporated into chromatin is constant, we compared relative abundances of differentially modified histone forms to each other. We highlighted the accumulation of highly acetylated histone forms during the SB-mediated inhibition of HDAC. Most of these variables were also identified as discriminant for the 2.5 mM SB dose, but with larger fold changes as summarized in Table 2. These results illustrate that our global differential approach was sensitive enough to emphasize a dose-response relationship. Moreover, it has been able to highlight modified forms of histones that accounted for only 0.1% of relative abundance.

Interestingly, modified discriminant forms of histone H3 were brought to light only for the 2.5 mM SB dose, meaning that H3 is less responsive to SB treatment than other type of core histones. More generally, 10 discriminant histone forms were identified in both groups, 7 were exclusively identified in the SB 1 mM group and 10 in the SB 2.5 mM one (Fig. 8).

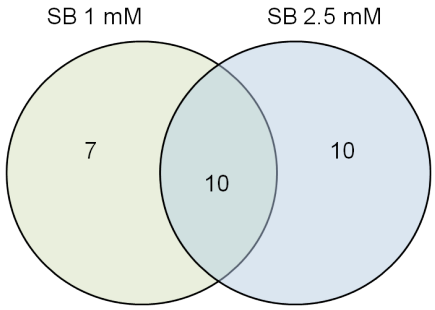


Fig. 8: Venn diagram of the discriminant histone forms obtained from the comparison of the two SB doses.

Regardless of the dose, the most significant effect of SB treatment was observed on histone H4, with 10 differently modified forms identified for the 1 mM SB dose and 7 for 2.5 mM SB dose. The 7 upregulated forms were 2 to 6 times acetylated. Among them, 3 monoacetylated forms were found as downregulated by SB treatment: the monoacetylated (11,277.5 Da) and the monoacetyl-monomethylated (11,291.5 Da) forms with both 1 mM and 2.5 mM SB doses, and the monoacetyl-dimethylated (11,305.5 Da) form with 1 mM SB dose. According to Pesavento *et al.*,[56] acetylation occurs predominantly on H4 dimethylated at the lysine-20 residue (K20me2 form) in a wide variety of human cancer cell lines, the most abundant H4 histone form being the S1Ac-K20me2 one. Close examination of the deconvoluted spectrum of histone H4 (Fig. 9) showed indeed that the most abundant form of H4 in control seemed to be the monoacetyl-dimethyl one (11,305.5 Da) compatible with the S1Ac-K20me2 species.

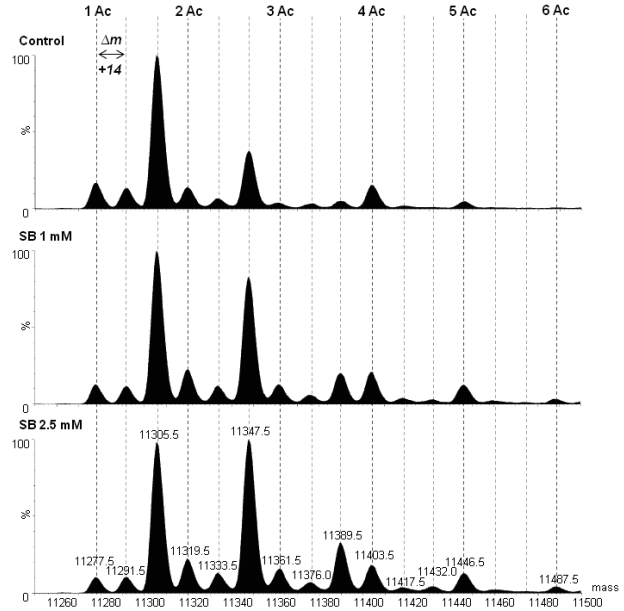


Fig. 9: Deconvoluted spectra of histone H4 in control (up), 1 mM (middle) and 2.5 mM (down) SB samples. Mass increments of +14 Da are highlighted by vertical dashed lines. In this case, a mass increment of +42 Da corresponds to an acetylation.

Four H3 forms were found to be less abundant following the 2.5 mM SB treatment while 3 appeared upregulated. The fact that the former had a shorter retention time during UPLC separation than the latter form (namely 11.5 *vs* 15.2 minutes) was consistent with their lower acetylation level, thus their lower relative hydrophobicity. According to peak relative abundance on deconvoluted spectra and measured molecular mass, we decided to argue on the H3.1 assumption. Among the less modified forms, the one with an increment of +28 Da was assigned to a H3 species with one dimethylated lysine, probably K9, rather than H3 with two methyl groups, on the basis of previous observations made on H3 from HeLa cells [57]. Similarly, species with a mass increment of either +56 or +70 Da were probably linked to H3K9me1-K14ac and H3K9me2-K14ac species. The +42 Da one was considered as the monoacetylated

Table 1: Most probable identifications of discriminant variables between control and 1 mM SB samples.
*Mean ratios below 1 have been recalculated as follows: $ratio = -1/(mean\,SB/mean\,Control)$

Putative Identity	UniProtKB Accession number	Theoretical average mass of unmodified form (Da)	Observed average mass of unmodified form (Da)	Observed average mass increment (Da)	PTMs	VIP Scores	%RSD	FDR (q-value)	Ratio*
H2A-1B/E	P04908	14,004.3	14,005	+42	1 ac	1.62	9.3	8.7E-08	1.38
H2B-1K	O60814	13,758.9	13,758.5	0	***	1.55	11.9	2.0E-07	-1.54
				+42	1 ac	1.65	16.7	6.1E-07	1.63
				+56	1 ac + 1 me1	2.07	4.9	2.0E-14	1.66
				+70	1 ac + 2 me1 / 1 me2	2.27	9.6	2.3E-09	1.77
				+84	2 ac	2.37	9.6	2.9E-07	1.87
H2B-1M	Q99879	13,858.0	13,857.5	+84	1 ac + 1 me2 + 1 me1	1.79	18	7.8E-05	1.60
H4	P62805	11,236.1	11,236.5	+42	1 ac	2.51	16.7	3.9E-14	-2.22
				+56	1 ac + 1me1	2.13	9	9.1E-13	-1.72
				+70	1 ac + 1me2	1.64	4.5	6.7E-14	-1.37
				+112	2 ac + 1me2	1.77	5.2	6.5E-13	1.48
				+126	3 ac	3.01	11.3	7.9E-09	3.10
				+154	3 ac + 1me2	3.08	7.7	8.1E-13	3.17
				+168	4 ac	2.84	7.6	9.4E-11	2.45
				+196	4 ac + 1me2	3.19	16.3	4.7E-14	3.35
				+210	5 ac	2.08	9.6	4.8E-09	1.74
				+252	6 ac	2.81	12.7	1.1E-07	2.59

Table 2: Most probable identifications of discriminant variables between control and 2.5 mM SB samples.
***Mean ratios below 1 have been recalculated as follows: ratio = −1/(mean SB/mean Control)**

Putative Identity	UniProtKB Accession number	Theoretical average mass of unmodified form (Da)	Observed average mass of unmodified form (Da)	Observed average mass increment (Da)	PTMs	VIP Scores	%RSD	FDR (q-value)	Ratio*
				+42	1 ac	1.82	14.5	1.6E-10	1.73
H2A-1B/E	P04908	14,004.3		+56	1 ac + 1 me1	1.80	19.5	2.8E-10	2.01
			14,005	+84	2 ac	1.81	27.3	2.3E-05	1.84
				+56	1 ac + 1 me1	1.79	5.0	2.5E-16	1.74
H2B-1K	O6081	13,758.9	13,758.5	+70	1 ac + 2 me1 / 1 me2	2.06	8.9	2.8E-15	2.12
				+84	2 ac	2.12	8.9	5.4E-11	2.19
				+28	1 me2	1.79	23.6	2.6E-06	-2.32
				+42	1 ac	1.83	16.9	1.5E-07	-2.61
				+56	1 ac + 1 me1	2.10	24.8	5.1E-07	-2.83
H3.1	P68431	15,272.9	15,272.5	+70	1 ac + 1 me2	2.06	19.0	2.6E-06	-2.96
				+98	2 ac + 1 me1	1.80	25.2	8.6E-04	2.01
				+112	2 ac + 2 me1 / 1 me2	1.76	29.1	8.9E-05	2.23
				+154	3 ac + 2 me1 / 1 me2	1.77	15.9	7.8E-05	2.18
				+42	1 ac	2.05	8.5	4.0E-17	-2.56
				+56	1 ac + 1me1	1.85	12.6	2.0E-15	-2.24
				+126	3 ac	2.45	15.4	3.3E-11	3.22
H4	P62805	11,236.1	11,236.5	+140	3 ac + 1me1	2.41	18.3	1.8E-09	3.00
				+154	3 ac + 1 me2	2.67	11.4	7.3E-16	3.94
				+168	4 ac	2.35	9.0	6.5E-15	2.78
				+196	4 ac + 1 me2	2.91	12.4	4.2E-18	4.92

form, modification occurring on a lysine residue but not at the N-terminus of H3.1.[57] As expected, these non- or monoacetylated species were decreased with the SB treatment while probable di- and triacetylated ones were increased. Assuming that acetylated residues could be K14, K23 in diacetylated species could be in good agreement with Thomas et al. findings.[57]
As an example, the relative abundance of each acetylation state for H4 according to the treatment is summarized on Fig. 10. These variations in relative amounts corroborate the results obtained by multivariate statistical approaches.

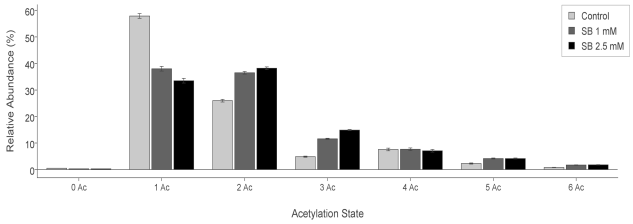


Fig. 10: Comparison of relative abundances for each acetylation state of histone H4 in control, 1 mM and 2.5 mM SB samples. Data represent mean ±SD for three biological replicates.

Material and methods

Cell culture and sodium butyrate treatment

Chemicals and cell culture reagents were respectively purchased from Sigma-Aldrich (Saint-Quentin-Fallavier, France) and Gibco (Cergy Pontoise, France). The BeWo cell line was originally derived from a human choriocarcinoma,[58] and the b30 clone was first obtained in Dr. Alan Schwartz laboratory (Washington University, St. Louis, MO, USA). This clone has the particularity to produce confluent monolayers within 4 to 6 days.[59] Cells were maintained at 37°C under 5% CO_2 atmosphere in Ham's F-12K (Kaighn's) medium supplemented with 10% fetal calf serum, 50 IU.mL^{-1} penicillin and 50 IU.mL^{-1} streptomycin. Subconfluent monolayers (n=15 per condition) were treated in 75 cm² flasks (Corning, NY, USA) with either 1 mM or 2.5 mM sodium butyrate dissolved in fresh F-12K medium. These doses were intentionally chosen to prevent cell cycle arrest caused by higher butyrate concentration (> 5 mM). Untreated cells were used as controls. After 24h, cells were individualized for approximately 5 min at 37°C with an adequate volume of Trypsin-EDTA (Eurobio, Courtaboeuf, France) and resuspended in fresh growth medium before cell counting (≈ 6 million cells per flask). After centrifugation, cell pellets were washed twice with ice-cold phosphate buffer saline (PBS) and dried before storage at -80°C.

Histone purification

Chemicals were either purchased from Sigma-Aldrich (Saint-Quentin-Fallavier, France), Roche (Meylan, France) for Nonidet P40, Complete ultra mini protease inhibitors cocktail and PhosStop, Bio-Rad (Marnes-la-Coquette, France) for electrophoresis or LaserBio Labs (Sophia-Antipolis, France) for alpha-cyano-hydroxycinnamic acid. BeWo histones were purified as described below. Briefly, each dried cell pellet was homogenised in lysis buffer (20 mM Hepes pH 7.5, 0.5 mM DTT, 1.5 mM $MgCl_2$, 10 mM KCl, 0.15% NP-40, 0.24 M sucrose, 10 mM sodium butyrate, protease and phosphatase inhibitors) and incubated on ice for 30 minutes with gentle stirring. Intact nuclei were pelleted by a 10 minutes centrifugation at 3,000 g and washed twice with cold lysis buffer without NP-40. Nuclei were then lysed with the same buffer as before containing only 0.05% NP-40, 2 mM additional EDTA and 0.3 M KCl added drop by drop in order to get rid of the non-histone proteins loosely bound with DNA. Chromatin pellet was recovered by a 15 minutes centrifugation at 16,000 g. Histones were extracted with 0.4 N H_2SO_4 for 4 hours at 4°C, and the acid soluble histones were recovered by centrifugation (20 minutes at 16,000 g). A 25% trichloroacetic acid (TCA) precipitation of histones was then done overnight at 4°C and the subsequent pellet washed three times with cold acetone.[60] The total histone dried pellet was finally resuspended in 20 mM Hepes pH 7.5. Total protein concentrations were determined by BCA assay (Thermo scientific, Villebon sur Yvette, France). Extraction efficiency was checked for each extract by both 13 % SDS-PAGE [61] stained with colloidal Coomassie Blue (Imperial™ Protein Stain, Thermo scientific, Villebon sur Yvette, France) and MALDI-MS analysis with a Voyager-DE™ Pro (linear positive ion mode, premix 1:1 with a 10 mg.mL^{-1} alpha-cyano-hydroxycinnamic acid solution in 0.3 % TFA in 50 % acetonitrile) from Applied Biosystems (Courtaboeuf, France).

UPLC-MS analysis of unfractionated intact histones

For the analysis, LC-MS grade solvents were purchased from J.T.Baker (Paris, France). All LC-MS experiments were performed on a Waters ACQUITY UPLC system (Waters, Milford, MA, USA) coupled to a hybrid quadrupole orthogonal time-of-flight mass spectrometer (SYNAPT G2 HDMS, Waters MS Technologies, Manchester, UK) operated in the positive electrospray ionization mode (ESI+). Orthogonal sequencing of the sample list was implemented to minimize the influence of injection order on the results.[62] In order to condition the column and later assess data quality and stability of the instrument, a pooled quality control sample (QC) was prepared by aliquoting identical volumes from each diluted study samples. This QC sample was injected ten times prior to running samples and then every 5 samples within the sample list. Histone extracts were

diluted to 0.3 $\mu g.\mu L^{-1}$ using the mobile phase A (see below) and an aliquot of 5 µL was loaded onto a Waters ACQUITY UPLC BEH C_{18} column (2.1 mm x 150 mm, 300 Å, 1.7 µm). Mobile phases were 0.05 % formic acid in water (solvent A) and 0.05 % formic acid in acetonitrile (solvent B). Chromatographic separation was performed at 55°C with a 0.3 $mL.min^{-1}$ constant flow rate using a linear gradient from 15 to 60 % B in 22 min, followed by 8 min re-equilibration at 15 % B. Samples were kept at 4°C in the autosampler during the analysis. Optimized mass spectrometer parameters were as follows: capillary voltage 3000 V, cone voltage 30 V, source temperature 120°C, desolvatation temperature 600°C, cone gas flow 20 $L.h^{-1}$, desolvatation gas flow 900 $L.h^{-1}$. The mass spectrometer was operated in the resolution mode with a scan time of 0.5 s. Data were collected in both centroid and continuum modes over the mass range *m/z* 50-2,000 with a resolution of 20,000. Mass measurement accuracy was ensured both by an initial sodium iodide calibration (2 $\mu g.\mu L^{-1}$ in isopropanol:water, 50:50, v/v) and by an intermittent flowing of leucine enkephalin (2 $ng.\mu L^{-1}$ in acetonitrile:water 50:50, v/v) at 15 $\mu L.min^{-1}$ as a lock mass compound. Data acquisition was carried out using Waters MassLynx™ v4.1 software (Waters MS Technologies, Manchester, UK). MaxEnt I deconvolution software (Waters MS Technologies, Manchester, UK) was used for generating deconvoluted spectra from acquired continuum spectra.

Data preprocessing

UPLC-MS centroid data files were converted to mzData format using a method previously described.[63] Briefly, raw files were directly read into the R software environment to extract information about the lock mass scans and the type of data, prior to being converted using massWolf program. The three-dimensional mzData files (*m/z*, retention time and ion current) were preprocessed using the freely available XCMS package[64] for the R statistical programming language. Peak detection and peak matching across samples were performed using the centWave algorithm.[65] Retention time (t_R) correction and chromatographic alignment were performed using the OBI-warp method,[66] before applying peak-picking on the integrated XICs (Extracted Ion Chromatograms). The preprocessing step results in an *X*-matrix where t_R and *m/z* values were concatenated into "t_R_*m/z*" features (in columns) present in each sample (in rows) with corresponding peak areas. To improve the performance of the statistical analysis and remove sources of systematic variation, the data were median normalised, log transformed and Pareto scaled. All normalisation steps were performed using in-house R scripts.

Multivariate statistical analysis

Prior to multivariate analysis, the complete dataset (*n*=45) was randomly split into a training set (*n*=10 for each class) for modeling and a blind test set containing the remaining samples (*n*=5 for each class) for validation purposes. Hierarchical cluster analysis (HCA) was performed by means of Ward's linkage as a clustering algorithm and Euclidean distance as a measure of dissimilarity between samples. The *X*-matrix was then exported to SIMCA-P13 software (MKS Umetrics, Umeå, Sweden) for chemometric analysis. Principal component analysis (PCA) was first applied to get an overview of the data and identify potential outliers. A supervised method (PLS-DA, Partial Least Squares-Discriminant Analysis) was used for classifying all samples according to the treatment conditions (control, 1 mM or 2.5 mM SB). A permutation test (999 iterations) was performed to prevent PLS-DA model from over fitting by comparing diagnostic statistics of the computed model with those of randomly generated models.[67] In the case of two-class discrimination, a significant simplification of both model interpretation and information extraction was obtained through OPLS-DA (Orthogonal Projections to Latent Structures-Discriminant Analysis).[68] OPLS-DA separates PLS components into one predictive component which concentrates all the discriminatory information, and one or more orthogonal (non-predictive) components that are unrelated to class discrimination.[69]
Cross-validated analysis of variance (CV-ANOVA) was systematically performed based on the cross-validated model.[70] We set the p-value threshold at 0.001 for indicating a significant group separation. External test set was then used to effectively test the predictive and discriminatory power of each generated models.[71] Once OPLS-DA models have being validated, variables with a VIP (Variable Importance in Projection) score beyond 1.5 on the predictive component were considered as contributing to the models and selected for further exploration.[72] Relative standard deviations (%RSD) for each variable in the QC replicates (*n*=19) were expressed as percentages and calculated as follows:

$$\%RSD \quad 100 \times \frac{\sigma}{\mu}$$

(σ stands for the standard deviation and μ for the mean)

All the variables with a %RSD value over 30% were discarded prior to univariate statistical analysis.[73]

Univariate statistical analysis

Subsequently, Welch's t-test for unequal variances was applied to each selected variable in order to confirm their actual difference between the two groups. The Benjamini and Hochberg false discovery rate (FDR) method[74] was used for calculating the false-positive rate associated with multiple comparisons, and

provides corrected q-values with a 0.01 significance level (FDR < 1%). Finally, mean ratios were computed for each variable to compare the absolute value change between treated and control samples.

Conclusion

Most of the current MS-based strategies rely on ultra-high resolution instruments and target specific PTMs sites at the protein (top-down) or the peptide (bottom-up and middle-down) level instead of focusing on the whole histone code changes. We have developed a new label-free differential proteomic approach capable of linking histone modification patterns to a specific xenobiotic exposure in a single LC-MS experiment. One of its main advantages, besides the straightforwardness and the small amount of proteins required, is to provide valuable information at the intact protein level without using ultra-high resolution capabilities or suitable type of fragmentation. Moreover, the employed multivariate techniques took into account how variables combine together to form discriminant patterns, which would not have been possible with conventional univariate statistics. In the current study, the sodium butyrate treatment illustrated the effectiveness of our approach for the relative quantification of histone modified forms. Unsupervised methods revealed that SB-mediated HDAC inhibition induced substantial changes in histone PTMs patterns. For each SB dose, supervised classification models succeeded in discriminating control from treated samples with 100% accuracy. Discriminant features were attributed to different acetylated and/or methylated histone forms. Though most of the selected histone markers were successfully identified, some ambiguities remained especially regarding the H2B histone forms that might be better resolved with adequate targeted strategies. This new global histonomic approach could prove to be a complementary tool for screening and monitoring changes in histone code occurring under toxicological or pathological conditions.

Acknowledgements

This study was supported by ANR PLACENTOX (ANR-11-CESA-0002). The PhD position of R.B. was funded by Région-Île-de-France (DIM SEnT). O.L. is indebted to Fondation pour la Recherche Médicale, Région-Île-de-France, Centre National de la Recherche Scientifique and ADRAPHARM Association for their financial support.
Special thanks to Abdellah Tebani for fruitful discussions on statistical analysis.

Notes and references

[a] UMR CNRS 8638, Faculté des Sciences Pharmaceutiques et Biologiques de Paris, Université Paris Descartes, Sorbonne Paris Cité, 4 avenue de l'Observatoire, 75006 Paris, France
[b] INSERM U1139, Faculté des Sciences Pharmaceutiques et Biologiques de Paris, Université Paris Descartes, Sorbonne Paris Cité, 4 avenue de l'Observatoire, 75006 Paris, France
[c] AP-HP, service de Toxicologie Biologique, Hôpital Lariboisière, 4 rue Ambroise Paré, 75475 Paris cedex 10, France
*corresponding author: sylvie.gillet@parisdescartes.fr, tel: (+33)1 53 73 99 32

Electronic Supplementary Information (ESI) available: Supplementary information is available as separate document. See DOI: 10.1039/b000000x/

1. K. Luger and T. J. Richmond, *Curr. Opin. Genet. Dev.*, 1998, **8**, 140–146.
2. T. Jenuwein and C. D. Allis, *Science*, 2001, **293**, 1074–1080.
3. K. A. Gelato and W. Fischle, *Biol. Chem.*, 2008, **389**, 353–363.
4. A. J. Bannister and T. Kouzarides, *Cell Res.*, 2011, **21**, 381–395.
5. A. Eberharter and P. B. Becker, *EMBO Rep.*, 2002, **3**, 224–229.
6. M. D. Shahbazian and M. Grunstein, *Annu. Rev. Biochem.*, 2007, **76**, 75–100.
7. B. D. Strahl and C. D. Allis, *Nature*, 2000, **403**, 41–45.
8. C. A. Musselman, M.-E. Lalonde, J. Côté, and T. G. Kutateladze, *Nat. Struct. Mol. Biol.*, 2012, **19**, 1218–1227.
9. R. Margueron, P. Trojer, and D. Reinberg, *Curr. Opin. Genet. Dev.*, 2005, **15**, 163–176.
10. S. Dik, P. T. J. Scheepers, and L. Godderis, *Crit. Rev. Toxicol.*, 2012, **42**, 491–500.
11. S. M. Fuchs, K. Krajewski, R. W. Baker, V. L. Miller, and B. D. Strahl, *Curr. Biol.*, 2011, **21**, 53–58.
12. A. Moradian, A. Kalli, M. J. Sweredoski, and S. Hess, *Proteomics*, 2014, **14**, 489–497.
13. A. L. Burlingame, X. Zhang, and R. J. Chalkley, *Methods*, 2005, **36**, 383–394.
14. B. A. Garcia, S. Mollah, B. M. Ueberheide, S. A. Busby, T. L. Muratore, J. Shabanowitz, and D. F. Hunt, *Nat. Protoc.*, 2007, **2**, 933–938.
15. R. Liao, H. Wu, H. Deng, Y. Yu, M. Hu, H. Zhai, P. Yang, S. Zhou, and W. Yi, *Anal. Chem.*, 2013, **85**, 2253–2259.
16. M. T. Boyne 2nd, J. J. Pesavento, C. A. Mizzen, and N. L. Kelleher, *J. Proteome Res.*, 2006, **5**, 248–253.
17. N. Siuti and N. L. Kelleher, *Nat. Methods*, 2007, **4**, 817–821.
18. Z. Tian, N. Tolić, R. Zhao, R. J. Moore, S. M. Hengel, E. W. Robinson, D. L. Stenoien, S. Wu, R. D. Smith, and L. Paša-Tolić, *Genome Biol.*, 2012, **13**, R86.

19. R. A. Zubarev, D. M. Horn, E. K. Fridriksson, N. L. Kelleher, N. A. Kruger, M. A. Lewis, B. K. Carpenter, and F. W. McLafferty, *Anal. Chem.*, 2000, **72**, 563–573.
20. J. E. P. Syka, J. J. Coon, M. J. Schroeder, J. Shabanowitz, and D. F. Hunt, *Proc. Natl. Acad. Sci. U.S.A.*, 2004, **101**, 9528–9533.
21. A. Kalli, M. J. Sweredoski, and S. Hess, *Anal. Chem.*, 2013, **85**, 3501–3507.
22. K. Contrepois, E. Ezan, C. Mann, and F. Fenaille, *J. Proteome Res.*, 2010, **9**, 5501–5509.
23. C. Zhang, Y. Liu, and P. C. Andrews, *Methods*, 2013, **61**, 236–243.
24. X. Guan, N. Rastogi, M. R. Parthun, and M. A. Freitas, *Mol. Cell Proteomics*, 2013, **12**, 2048–2059.
25. L. Adwan and N. H. Zawia, *Pharmacol. Ther.*, 2013, **139**, 41–50.
26. A. Fischer, *Neuropharmacology*, 2014.
27. P. Chaturvedi and S. C. Tyagi, *Int. J. Cardiol.*, 2014.
28. N. Takai and H. Narahara, *Cancers*, 2010, **2**, 1683–1688.
29. J. E. Bolden, M. J. Peart, and R. W. Johnstone, *Nat. Rev. Drug Discov.*, 2006, **5**, 769–784.
30. K. Zhang, M. Schrag, A. Crofton, R. Trivedi, H. Vinters, and W. Kirsch, *Proteomics*, 2012, **12**, 1261–1268.
31. K. L. Fiedler, P. Bheda, J. Dai, J. D. Boeke, C. Wolberger, and R. J. Cotter, *J. Mass Spectrom.*, 2013, **48**, 608–615.
32. W. Weckwerth and K. Morgenthal, *Drug Discov. Today*, 2005, **10**, 1551–1558.
33. A. K. Jain, R. P. W. Duin, and J. Mao, *IEEE Transactions on pattern analysis and machine intelligence*, 2000, **22**, 4 – 37.
34. H. Spratt, H. Ju, and A. R. Brasier, *Methods*, 2013, **61**, 73–85.
35. J. A. Kirwan, D. I. Broadhurst, R. L. Davidson, and M. R. Viant, *Anal. Bioanal. Chem.*, 2013, **405**, 5147–5157.
36. K. A. Veselkov, L. K. Vingara, P. Masson, S. L. Robinette, E. Want, J. V. Li, R. H. Barton, C. Boursier-Neyret, B. Walther, T. M. Ebbels, I. Pelczer, E. Holmes, J. C. Lindon, and J. K. Nicholson, *Anal. Chem.*, 2011, **83**, 5864–72.
37. A. C. Sauve and T. P. Speed, in *In Procedings Gensips*, 2004.
38. L. Eriksson, J. Trygg, and S. Wold, *J. Chemom.*, 2014, **28**, 332–346.
39. A. M. De Livera, D. A. Dias, D. De Souza, T. Rupasinghe, J. Pyke, D. Tull, U. Roessner, M. McConville, and T. P. Speed, *Anal. Chem.*, 2012, **84**, 10768–76.
40. R. A. van den Berg, H. C. Hoefsloot, J. A. Westerhuis, A. K. Smilde, and M. J. van der Werf, *BMC Genomics*, 2006, **7**, 142.
41. L. Eriksson, H. Antti, J. Gottfries, E. Holmes, E. Johansson, F. Lindgren, I. Long, T. Lundstedt, J. Trygg, and S. Wold, *Anal. Bioanal. Chem.*, 2004, **380**, 419–429.
42. J. Xia, N. Psychogios, N. Young, and D. S. Wishart, *Nucleic Acids Res.*, 2009, **37**, W652–W660.
43. S. Wold, *Technometrics*, 1978, **20**, 198–200.
44. H. Albazzaz and X. Z. Wang, *Ind. Eng. Chem. Res.*, 2004, **43**, 6731–6741.
45. Z. Huang, L. Lin, Y. Gao, Y. Chen, X. Yan, J. Xing, and W. Hang, *Mol. Cell Proteomics*, 2011, **10**, 1–10.
46. F. Lindgren, B. Hansen, W. Karcher, M. Sjostrom, and L. Eriksson, *J. Chemom.*, 1998, **10**, 521–532.
47. V. Esposito Vinzi, W. . Chin, J. Henseler, and H. Wang, *Handbook of Partial Least Squares*, Springer., 2010.
48. Å. M. Wheelock and C. E. Wheelock, *Mol. Biosyst.*, 2013, **9**, 2589–2596.
49. A. G. Ferrige, M. J. Seddon, S. Jarvis, J. Skilling, and R. Aplin, *Rapid Commun. Mass Spectrom.*, 1991, **5**, 374–377.
50. E. Gasteiger, C. Hoogland, A. Gattiker, S. Duvaud, M. R. Wilkins, R. D. Appel, and A. Bairoch, in *The Proteomics Protocols Handbook*, ed. J. M. Walker, Humana Press, 2005, pp. 571–607.
51. S. P. Khare, F. Habib, R. Sharma, N. Gadewal, S. Gupta, and S. Galande, *Nucleic Acids Res.*, 2012, **40**, D337–D342.
52. The UniProt Consortium, *Nucleic Acids Research*, 2013, **42**, D191–D198.
53. I. Ntai, K. Kim, R. T. Fellers, O. S. Skinner, A. D. Smith, B. P. Early, J. P. Savaryn, R. D. LeDuc, P. M. Thomas, and N. L. Kelleher, *Anal. Chem.*, 2014, **86**, 4961–4968.
54. X. Su, D. M. Lucas, L. Zhang, H. Xu, V. Zabrouskov, M. E. Davis, A. R. Knapp, D. C. Young, P. R. O. Payne, M. R. Parthun, G. Marcucci, M. R. Grever, J. C. Byrd, and M. A. Freitas, *Proteomics*, 2009, **9**, 1197–1206.
55. T. Jiang, X. Zhou, K. Taghizadeh, M. Dong, and P. C. Dedon, *Proc. Natl. Acad. Sci. U.S.A.*, 2007, **104**, 60–65.
56. J. J. Pesavento, H. Yang, N. L. Kelleher, and C. A. Mizzen, *Mol. Cell. Biol.*, 2008, **28**, 468–486.
57. C. E. Thomas, N. L. Kelleher, and C. A. Mizzen, *J. Proteome Res.*, 2006, **5**, 240–247.
58. R. A. Pattillo and G. O. Gey, *Cancer Res.*, 1968, **28**, 1231–6.
59. F. Liu, M. J. Soares, and K. L. Audus, *Am. J. Physiol.*, 1997, **273**, C1596–604.
60. D. Shechter, H. L. Dormann, C. D. Allis, and S. B. Hake, *Nat. Protoc.*, 2007, **2**, 1445–1457.
61. U. K. Laemmli, *Nature*, 1970, **227**, 680–685.
62. V. Croixmarie, T. Umbdenstock, O. Cloarec, A. Moreau, J. M. Pascussi, C. Boursier-Neyret, and B. Walther, *Anal. Chem.*, 2009, **81**, 6061–9.
63. J. Stanstrup, M. Gerlich, L. O. Dragsted, and S. Neumann, *Anal. Bioanal. Chem.*, 2013, **405**, 5037–48.
64. C. A. Smith, E. J. Want, G. O'Maille, R. Abagyan, and G. Siuzdak, *Anal. Chem.*, 2006, **78**, 779–87.
65. R. Tautenhahn, C. Bottcher, and S. Neumann, *BMC Bioinformatics*, 2008, **9**, 504.
66. J. T. Prince and E. M. Marcotte, *Anal. Chem.*, 2006, **78**, 6140–52.

67. E. Szymanska, E. Saccenti, A. K. Smilde, and J. A. Westerhuis, *Metabolomics*, **8**, 3–16.
68. J. M. Fonville, S. E. Richards, R. H. Barton, C. L. Boulange, T. M. D. Ebbels, J. K. Nicholson, E. Holmes, and M.-E. Dumas, *J. Chemom.*, 2010, **24**, 636–649.
69. J. Trygg and S. Wold, *J. Chemom.*, 2002, **16**, 119–128.
70. S. Eriksson, L., Trygg, J., Wold, *J. Chemom.*, 2008, **22**, 594–600.
71. K. H. Esbensen and P. Geladi, *J. Chemom.*, 2010, **24**, 168–187.
72. F. Marini, A. Roncaglioni, and M. Novic, *J. Chem. Inf. Model.*, 2005, **45**, 1507–19.
73. J. Sun, Y. Ando, D. Ahlbory-Dieker, L. K. Schnackenberg, X. Yang, J. Greenhaw, L. Pence, F. Qian, W. F. Salminen, and D. L. Mendrick, *J. Mol. Biomark. Diagn. S.*, 2013, **1**, 2.
74. Y. Benjamini and Y. Hochberg, *Journal of the Royal Statistical Society. Series B (Methodological)*, 1995, **57**, 289–300.