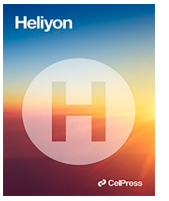
Heliyon 10 (2024) e29936

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Intact cell mass spectrometry coupled with machine learning reveals minute changes induced by single gene silencing

Lukáš Pečinka [a,b], Lukáš Moráň [c,d], Petra Kovačovicová [b,c], Francesca Meloni [e], Josef Havel [a,b], Tiziana Pivetta [e], Petr Vaňhara [b,c,*]

[a] Department of Chemistry, Faculty of Science, Masaryk University, Brno, Czech Republic
[b] International Clinical Research Center, St. Anne's University Hospital Brno, Czech Republic
[c] Department of Histology and Embryology, Faculty of Medicine, Masaryk University, Brno, Czech Republic
[d] Research Centre for Applied Molecular Oncology (RECAMO), Masaryk Memorial Cancer Institute, Brno, Czech Republic
[e] Chemical and Geological Sciences Department, University of Cagliari, Cittadella Universitaria, Monserrato, Italy

ARTICLE INFO

Keywords:
Intact cell MALDI TOF MS
Machine learning
Biotyping
TUSC3
R programming language
Bioinformatics
Quality control
Cell culture

ABSTRACT

Intact (whole) cell MALDI TOF mass spectrometry is a commonly used tool in clinical microbiology for several decades. Recently it was introduced to analysis of eukaryotic cells, including cancer and stem cells. Besides targeted metabolomic and proteomic applications, the intact cell MALDI TOF mass spectrometry provides a sufficient sensitivity and specificity to discriminate cell types, isogenous cell lines or even the metabolic states. This makes the intact cell MALDI TOF mass spectrometry a promising tool for quality control in advanced cell cultures with a potential to reveal batch-to-batch variation, aberrant clones, or unwanted shifts in cell phenotype. However, cellular alterations induced by change in expression of a single gene has not been addressed by intact cell mass spectrometry yet. In this work we used a well-characterized human ovarian cancer cell line SKOV3 with silenced expression of a tumor suppressor candidate 3 gene (TUSC3). TUSC3 is involved in co-translational N-glycosylation of proteins with well-known global impact on cell phenotype. Altogether, this experimental design represents a highly suitable model for optimization of intact cell mass spectrometry and analysis of spectral data. Here we investigated five machine learning algorithms (k-nearest neighbors, decision tree, random forest, partial least squares discrimination, and artificial neural network) and optimized their performance either in pure populations or in two-component mixtures composed of cells with normal or silenced expression of TUSC3. All five algorithms reached accuracy over 90 % and were able to reveal even subtle changes in mass spectra corresponding to alterations of TUSC3 expression. In summary, we demonstrate that spectral fingerprints generated by intact cell MALDI-TOF mass spectrometry coupled to a machine learning classifier can reveal minute changes induced by alteration of a single gene, and therefore contribute to the portfolio of quality control applications in routine cell and tissue cultures.

1. Introduction

Matrix-assisted laser desorption/ionization time of flight mass spectrometry (MALDI-TOF MS) is a widely used analytical technique

* Corresponding author. Masaryk University, Faculty of Medicine, Kamenice 5, Brno, Czech Republic.
E-mail address: PVanhara@med.muni.cz (P. Vaňhara).

https://doi.org/10.1016/j.heliyon.2024.e29936
Received 12 April 2024; Accepted 17 April 2024
Available online 22 April 2024

for identification, structural analysis, and quantification of various chemical species, including those constituting complex biological samples. In the last decades, MALDI TOF MS found its way even beyond analytical chemistry and has been introduced into clinical microbiology, histology, and cell biology [1,2]. The intact cell MALDI TOF MS uses the whole, undisturbed cells as input analytes, without the preceding cell lysis, fractionation, or protein extraction. In clinical microbiology, the mass spectra generated by MALDI TOF MS of intact bacteria provide unique cellular fingerprints specific for the biotyping of bacterial species and strains, and therefore facilitate the diagnostics without the need of extensive bacterial culture [3]. However, the biotyping of eukaryotic cells may represent a technological limitation due to the inherent complexity of samples, requiring the development and/or optimization of protocols tailored to individual experiments or applications. This involves the standardization of the protocol, in particular, manipulation and handling of sample (cell harvesting and washing), preparation of sample for MS (matrix choice, solvent composition, additives, sample spotting), and instrumental setup (type of instrument, laser wavelength, and *m/z* range and acquisition parameters). The mathematical methods for spectra pre-processing and data evaluation including multivariate statistical analysis are then determining for proper data interpretation. The principal component analysis (PCA), partial least square-discrimination analysis (PLS-DA), hierarchical clustering analysis (HCA), and machine learning (ML) are fast-growing approaches in MS data analysis and cell biotyping in particular [4,5]. Altogether, the intact cell MALDI TOF MS coupled with suitable mathematical apparatus represents a low-cost, robust simple, and straightforward method that allows discrimination of individual cell types, either alone or in mixtures or even the individual cell states during differentiation or acquisition of abnormal phenotypes [6,7].

Currently, several biotyping studies demonstrated discrimination of individual cell types based on informative regions in mass spectra. Karger et al. published an extensive study of 66 cell lines from 34 species ranging from insects to primates, that have been correctly classified by the mass spectra [8]. Two different pancreatic hormone-secreting cell lines were distinguished by Buchanan et al. [9]. Kober et al. demonstrate the ability to differentiate toxic effects in cell-based ecotoxicological test systems [10]. In 2015, in our study by Valletta et al. reported that MS combined with artificial neural networks (ANN) can quantitatively estimate cell numbers in binary mixtures of mouse embryonic stem cells and mouse fibroblasts, or mouse and human embryonic stem cells [6]. In a similar study published later, Petukhova et al. demonstrated discrimination of ovarian cancer cell lines and primary cells in two-component mixtures [11]. In 2018, we demonstrated that the intact cell MALDI TOF MS-based approach is also highly suitable for monitoring subtle alterations in phenotype of a single stem cell line over time, or during differentiation of stem cells and progenitors towards terminal phenotypes [12]. In summary, the intact cell MALDI-TOF MS combined with proper biostatistical tools can offer a versatile tool for quality control in preclinical or clinical-grade cell cultures, cancer cells discrimination, or gamete phenotyping [7,11,13,14].

Up to now, there is no dedicated publication reporting the use of MS biotyping for revealing alterations induced by a single gene expression change in a well-defined cellular model. We were therefore curious if the protocol we established previously for stem cell quality control can be used for tracing minute changes in cell phenotype [7]. As a model, we used an ovarian cancer cell line with downregulated gene coding for tumor suppressor candidate 3 (*TUSC3* or *N33*). The TUSC3 protein constitutes a subunit of the oligosaccharyltransferase subunit, contributing to the final steps of N-glycosylation of proteins in endoplasmic reticulum [15]. Loss of *TUSC3* expression alters the glycosylation of surface molecules and subsequently the proliferation, migration, and cell stress response in ovarian cancer cells, and other cancer cell lines [16–18]. In ovarian cancer patients, epigenetic loss of *TUSC3* expression correlates with poor prognosis and reduced survival [19]. We investigated the *TUSC3* gene previously, and prepared several well-characterized cellular models, therefore, for this study we chose the *TUSC3*-silenced SKOV3 ovarian cancer cell line. In the SKOV3 cells, expression of *TUSC3* gene was downregulated by short hairpin RNA (shRNA) as described previously [16,17]. In this work, the statistical analysis of the mass spectra clearly showed different spectral profiles in 2–20 kDa range in cells with normal and silenced *TUSC3*. Robustness of developed method was tested using mass spectra recorded in different technical replicates prepared separately on different days. Effect of normalization and number of *m/z* values as input was tested with the aim of how these data processing affect the accuracy of classification models. In summary, we demonstrated the intact cell MALDI-TOF MS when coupled to the proper statistical classifier can reveal changes induced by alteration of a single gene.

2. Material and methods

2.1. Cell culture

SKOV3 cells were obtained from American Type Culture Collection (USA) and cultured in high glucose (4.5 g/L) Dulbecco's Modified Eagle Medium (DMEM) enriched with 10 % Fetal Calf Serum (FCS) and 1 % Penicillin/Streptomycin sulphate, at 37 °C in humidified atmosphere containing 5 % of CO_2. TUSC3 expression was downregulated by short hairpin RNA (shRNA) encoded in the lentiviral plasmid pLKO.1 and transduced into SKOV3 cells as described previously [18]. pLKO.1 plasmid harboring scrambled short hairpin sequence was used as a control. Mycoplasma contamination was investigated on a routine basis using PCR. For analysis, cells were enzymatically detached, washed in phosphate buffered saline, counted, and stored in −80 °C as dry pellets.

2.2. RNA isolation, cDNA synthesis and quantitative real-time RT-PCR

Total RNA from SKOV3 cell line was isolated using the RNeasy Mini kit (Qiagen) and the quantity and purity were assessed by UV spectrometry at 260, 280 and 230 nm. cDNA was synthesized from 1 μg DNase I-digested total RNA using the First-strand cDNA Synthesis Kit (Sigma-Aldrich). Expression was relatively quantified using TaqMan probes specific for TUSC3, Hs00185147_m1 and β2-microglobulin, Hs99999907_m1 (Applied Biosystems) as described elsewhere and expressed as expression fold change [18]. All PCR reactions were performed from at least three independent experiments, and reverse transcriptase-negative and template-negative

controls were included.

2.3. Western blotting and immunoprecipitation

Harvested cells were washed two times with 1 × PBS and resuspended in the NP-40 lysis buffer containing 50 mM Tris-Cl (pH 7.4), 150 mM NaCl, 2 mM EDTA, 1 % NP-40, 50 mM NaF and supplemented with phosphatase inhibitor cocktail (Sigma Aldrich) and protease inhibitor cocktail (Complete, Roche). Protein extracts (15 μg) quantified by BCA protein assay (Pierce, Austria), were mixed with 2 × Laemmli sample buffer (100 mM Tris pH 6.8, 4 % SDS, 200 mM DTT, 20 % glycerol and 0.1 % bromophenol blue) boiled for 3 min and resolved by 10 % sodium dodecylsulfate-polyacrylamide gel electrophoresis (SDS-PAGE). The resolved proteins were then electroblotted to the 0.45-μm PVDF membrane (Millipore) and incubated with the primary antibodies (TUSC3, Ab65213 and actin, Ab1801 both from Abcam, UK) and diluted 1:500–1:1,000 at 4 °C overnight. The blots were developed using horseradish peroxidase-conjugated secondary antibodies (anti-rabbit HRP no. 7074 (Cell Signaling, USA) anti-mouse HRP Ab50043 (Abcam, UK), both 1:4,000 and Immobilon Western HRP substrate (Millipore, Czech Republic) according to the manufacturer's instructions.

2.4. Chemicals for mass spectrometry

Sinapinic acid (SA), alpha-cyano-4-hydroxycinnamic acid (CHCA), trifluoroacetic acid (TFA), and ammonium bicarbonate (ABC) were purchased from Sigma-Aldrich (Steinheim, Germany). Acetonitrile (ACN) was purchased from Penta (Prague, Czech Republic). Water was double distilled using a quartz apparatus from Heraeus Quarzschmelze (Hanau, Germany). IVD bacterial test standard (BTS) was purchased from Bruker Daltonik GmbH (Bremen, Germany).

2.5. Sample preparation for intact cell MALDI TOF MS analysis

Matrix solution was prepared by dissolving 30 mg of SA in 1 mL of ACN/H_2O (70/30 v/v) acidified by 7.5 % of TFA. Frozen cell pellets were thawed on ice and diluted with ice-cold 150 mM ABC buffer. The cell suspension in ABC was then mixed with SA matrix to reach the final concentration 12.5 × 10^6 cells per mL. 2 μL of the suspension containing 25 × 10^3 cells were spotted on 384-well steel target plate in technical replicates (n = 5) and dried at room temperature under dust-free conditions.

2.6. Spectra acquisition and processing

For mass spectrometry, the MALDI 7090 TOF-TOF instrument (Shimadzu Kratos Analytical) equipped with the 2 kHz ultra-fast solid-state UV laser (Nd-YAG: 355 nm) was used. Mass spectra were acquired in the linear positive ion mode, in mass region 2–20 kDa, with pulse extraction set to 12.5 kDa. The calibration was performed externally using the Bacterial Test Standard 3.5–17 kDa. In total 175 mass spectra acquired from seven different mixtures of $SKOV3_{scrambled\ shRNA}$ and $SKOV3_{TUSC3\ shRNA}$ in five technical replicates recorded in five different days were used.

The mass spectra exported in the mzML file format were pre-processed using the R project (4.0.4), MALDIquant package, MALDIrppa, and subsequently analyzed using additional R packages enabling multivariate statistical modeling [20]. Before pre-processing of mass spectra, the low-quality spectra were identified by semi-automatic screening implemented in the MALDIrppa package. The procedure is based on robust scale estimators of median intensities and derivative spectra [21]. The spectral pre-processing workflow followed standard procedures adopted from the MALDIquant package. Standard processing of mass spectra involved several steps:

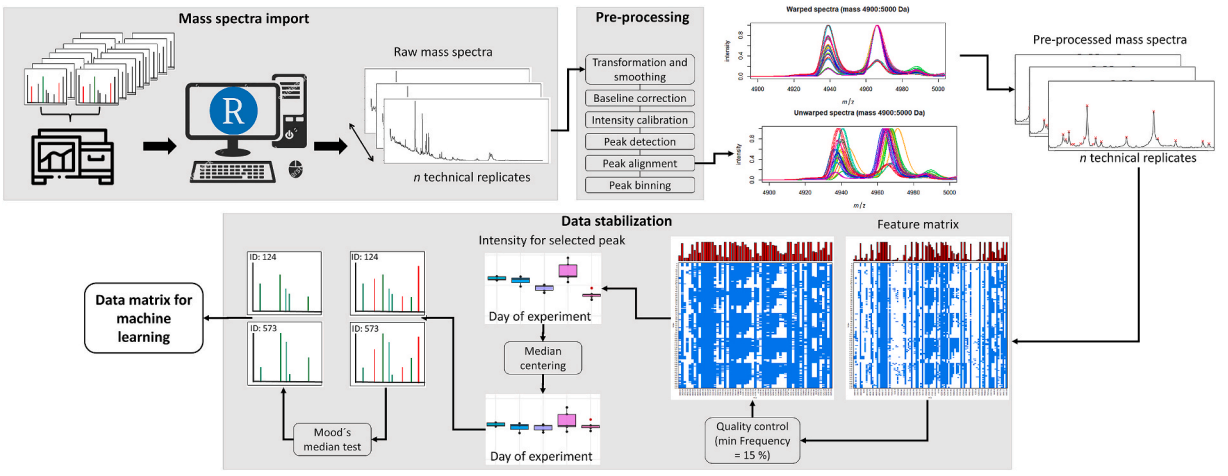


Fig. 1. Mass spectra processing workflow. Raw mass spectra in n technical replicates were processed, and the informative features selected. Data stabilization was performed to reduce the technical variability and unwanted signal bias as an inherent intra-experimental quality control.

briefly, the initial quality control, transformation and smoothing (Savitzky-Golay filter) with halfwindowSize = 100, baseline correction (statistics-sensitive non-linear iterative peak-clipping, SNIP) with 500 iterations, intensity calibration ($\sum X_i = 1$, where X_i are intensities of corresponding peaks in mass spectra), spectra alignment (removing the non-systematic shift in technical replication acquired on a different day), trimming (2–10 kDa), and peak detection using MAD noise estimation algorithm with signal-to-noise = 10 and a half-window size = 20 [22,23]. To eliminate artifacts in spectral data, the feature matrix of detected peaks was constructed only from the peaks that were detected at least in 20 % of total mass spectra. Peak lists for all mass spectra were converted to the feature matrix. Established matrix *m* x *n* consists of spectral data, where *m* represents selected *m/z* values and *n* are the IDs of individual samples. The *i-th* row of the matrix (*n*) shows the intensities of selected peaks (*m*) of the *i-th* samples (mixture). The feature matrix reduces the data from the original $n \times 400\,000$ to $n \times 75$. Established matrix of spectral data was used for further multivariate statistical methods and development of selected classifiers. In summary, the process is schematically described in Fig. 1.

2.7. Multivariate statistical analysis

A total of 175 mass spectra measured were analyzed in this study. PCA was initially used to reveal the data structure and to verify the results visually. The Partial Least Squares-Discriminant Analysis (PLS-DA), k-nearest Neighbors Algorithm (k-NN), Random Forest (RF), Decision Tree (DT), and ANN algorithms were chosen for the development of classification models. Data were split into training (70 %) and test cohorts (30 %). Cross-validation (CV) was performed. Considering the size of the sample, a 10 × repeated 5-fold CV was used with the "one standard error (SE)" rule for selecting the least complex model with the average cross-validated accuracy within the lowest root mean square error (RMSE) from that in the optimal model. All multivariate analyses and modeling were done in the R environment and validated in OriginPro 2023b SR1 licensed to the University of Cagliari.

3. Results and discussion

Short hairpin RNA-mediated silencing of *TUSC3* gene was performed as described previously by Vaňhara (2013) and validated by quantitative real time PCR and western blotting (Fig. 2) [17]. Expression of *TUSC3* by shRNA was downregulated to approximately 30 % when compared to control cells containing scrambled shRNA sequence.

For intact cell MALDI TOF MS, the samples of $SKOV3_{scrambled\ shRNA}$ and $SKOV3_{TUSC3\ shRNA}$ cells were prepared as described above and analyzed independently under identical conditions in five different days to document the technical reproducibility of the measurements. The representative mass spectra of $SKOV3_{scrambled\ shRNA}$ and $SKOV3_{TUSC3\ shRNA}$ cells are demonstrated in Fig. 3. In $SKOV3_{TUSC3\ shRNA}$ spectral data, several peaks show clear up- or downregulation, e.g. *m/z* 4910 ± 1 Da and 6087 ± 1 Da (downregulation) and 4937 ± 1 Da (upregulation) (Fig. S1). The Kendall correlation coefficient $\tau = 0.26$ for $SKOV3_{scrambled\ shRNA}$ and $SKOV3_{TUSC3\ shRNA}$ spectral datasets indicated low similarity between the groups. The $SKOV3_{scrambled\ shRNA}$ and $SKOV3_{TUSC3\ shRNA}$ groups, however, contain minor intrinsic variability, induced probably by pipetting errors or by inherent inconstancies during sample handling (spotting, sample plate cleaning, protein degradation in process of preparation, the residues of buffer within cell pellets), or the physical variability of the method itself (Fig. S2). The Kendall correlation coefficient within the datasets reached $\tau = 0.95$ for $SKOV3_{scrambled\ shRNA}$ and $\tau = 0.92$ for $SKOV3_{TUSC3\ shRNA}$, respectively.

The visual observation was confirmed by PCA and HCA. Both analyses clearly discriminated the $SKOV3_{scrambled\ shRNA}$ and $SKOV3_{TUSC3\ shRNA}$ cells. HCA was then performed to assess the relative hierarchical contribution of differences in molecular profiles in cell samples. PCA performed on data revealed separated clusters corresponding to $SKOV3_{scrambled\ shRNA}$ and $SKOV3_{TUSC3\ shRNA}$ cells (Fig. 4A), however, an obvious grouping effect inside the cell clusters related to batch effects of individual measurements was present.

This phenomenon was reported e.g. in mass spectrometry imaging, quantitative MS and proteomics or bacterial biotyping [24–26]. In eukaryotic cells, the batch effects of MS can decrease the sensitivity of analysis and limit the biotyping-based quality control [7]. The

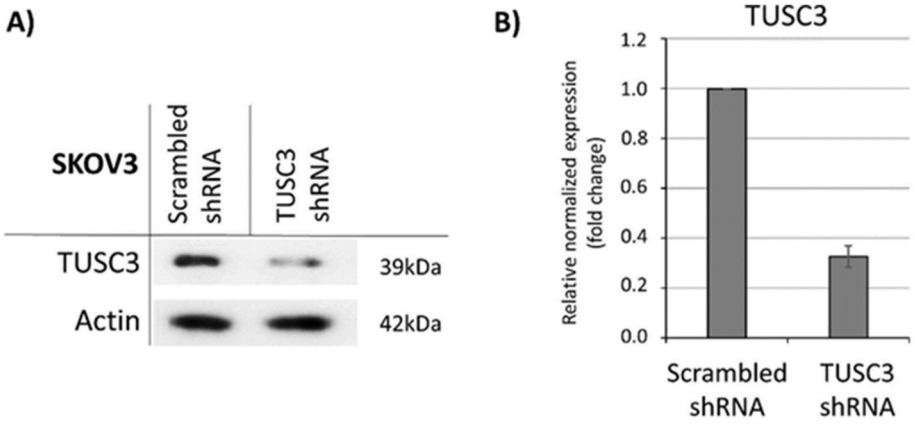


Fig. 2. Expression of *TUSC3* was downregulated by shRNA transduced into the SKOV3 cells. The decrease of *TUSC3* expression is visualized by western blotting (**A**), and by qRT-PCR as mean ± SD fold change (**B**).

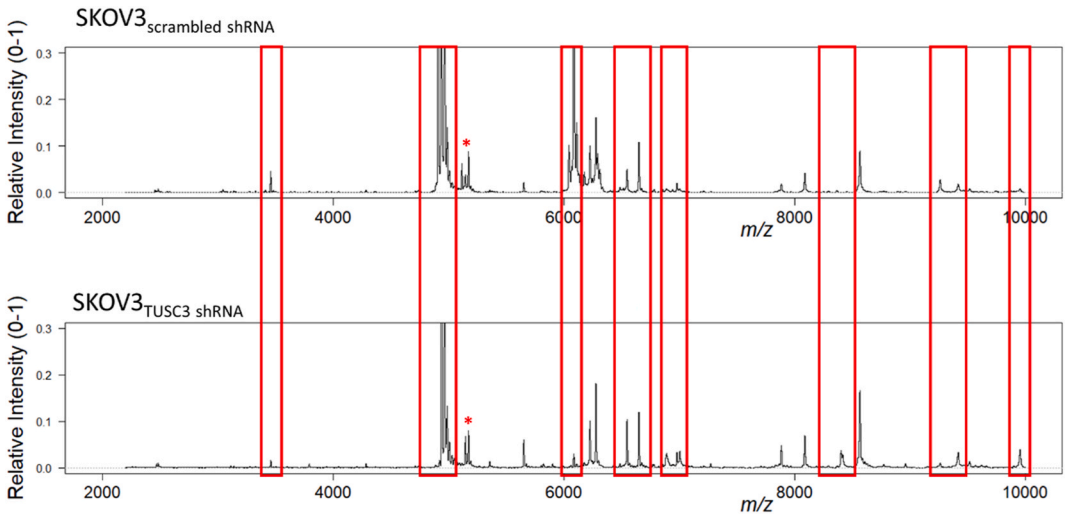


Fig. 3. Representative mass spectra of SKOV3$_{\text{scrambled shRNA}}$ and SKOV3$_{\text{TUSC3 shRNA}}$ cells. Highlighted regions indicate the *m/z* values that significantly differ in peak intensity. Asterisk (*) indicates the *m/z* values corresponding to the SA matrix adduct (+206 Da).

batch effect (Fig. 4A) is caused by pipetting errors or natural irregularities in sample handling (spotting, cleaning of sample plates, protein degradation during the preparation process, buffer residues in cell pellets) or the principial variability of the method itself (MALDI TOF MS). The phenomenon of batch effects makes the MALDI MS method generally not suitable for absolute quantification, and represents an issue also in intact cell mass spectrometry. The physico-chemical reasons of batch effects involve the process of sample-matrix crystallization, where the size of co-crystals and the effect of "sweet spots" cause variations in the measured intensities All of these discrepancies create an overall undesired variability in the measurement, which is usually referred to as the 'batch effect'. To overcome these limits and reduce the measurement-dependent fluctuations, the values of peak intensities were median centered (Fig. 4A and B) and Mood's median test (alpha = 0.05) was used to identify those peaks that significantly differed between groups [27]. Despite the fact we were able to acquire the peaks at the particular *m/z* values reproducibly, the event of repeated measurement introduced the unwanted variability into the peak intensities, as illustrated in Fig. 4C using 4939 Da peak intensity as an example. This explains the observation from PCA analysis. Fig. 4D demonstrates centering effect reducing fluctuations of 4939 Da peak intensity in SKOV3$_{\text{scrambled shRNA}}$ and SKOV3$_{\text{TUSC3 shRNA}}$ cells. To further visualize the similarities and divergences in the mass spectra, we constructed the heat maps based on peak intensities (Fig. 4E and F). All data in the spectral matrix were Z-score normalized across the groups. The heatmaps clearly revealed two clusters in data, dependent solely on peak intensities. Similarly, the reduction of the batch effect by median centering improved the discrimination of the SKOV3$_{\text{scrambled shRNA}}$ and SKOV3$_{\text{TUSC3 shRNA}}$ heat maps. The grouping effect observed in Fig. 4A was reduced upon median centering of data (Fig. 4B). First two principal components (PC1 and PC2) cover approximately 85 % of total data variability in both the non-centered and median-centered data. However, when the median-centered data were used, the contribution of PC1 increased, indicating a decrease of the technical variability in spectral data. In summary, we used the median-centered data for further analysis.

We were then curious whether we can discriminate spectral patterns of SKOV3$_{\text{TUSC3 shRNA}}$ cells in two-component mixtures with SKOV3$_{\text{scrambled shRNA}}$ control cells using the same panel of peaks as described above. We prepared analytes containing SKOV3$_{\text{scrambled shRNA}}$ and SKOV3$_{\text{TUSC3 shRNA}}$ cells in a series of volume ratios (1:0, 9:1, 8:2, 1:1, 2:8, 1:9, and 0:1) and then processed for mass spectrometry as described above (Fig. S3).

Computations involved two methodological approaches: 1) model-free, unsupervised, with light pre-treatment, and 2), model-free, unsupervised, and with light pre-treatment, and the second one with pre-treatment and more addressed to the differentiation of the cells. In the first method, raw mass spectra were resampled (resample factor 55) and transformed to the matrix consisted of 5070 rows (signals from mass spectra) and 175 columns (mass spectra). The data were then normalized by mean centering or by standard deviation division and analyzed by PCA (Fig. 5A and B). The corresponding loading plot then provided the importance of each variable for PC1, PC2, and PC3 (Fig. S4). The first method shows that each group of cell mixtures lies in a specific plane, indicating the possible classification.

The second method follows the procedure described in the section "Spectra acquisition and processing". Eigenvalue analysis indicated presence of three major factors, where two factors contribute to the overall variability ~80 %. The PCA analysis clearly discriminated the pure SKOV3$_{\text{scrambled shRNA}}$ and SKOV3$_{\text{TUSC3 shRNA}}$ as well as the respective cell mixtures (Fig. 6A). The relative abundance of five preselected *m/z* values is shown using normalized stacked bar chart in Fig. 6B. Signal at 4910 ± 1 Da decreased when *TUSC3* is silenced whereas the abundance of signal at 4937 ± 1 Da increased. Several *m/z* values were downregulated upon *TUSC3* silencing (e.g. 4910, 5117, 6045, and 6087 Da) and only a few were upregulated (e.g. 4937 and 8566 Da).

Unsupervised hierarchical cluster analysis using Euclidean distance matrix with Ward's method further confirms results obtained by PCA (Fig. 6C). The hierarchical cluster analysis identified groups based on the similarity of mass spectra fingerprints. Seven distinct

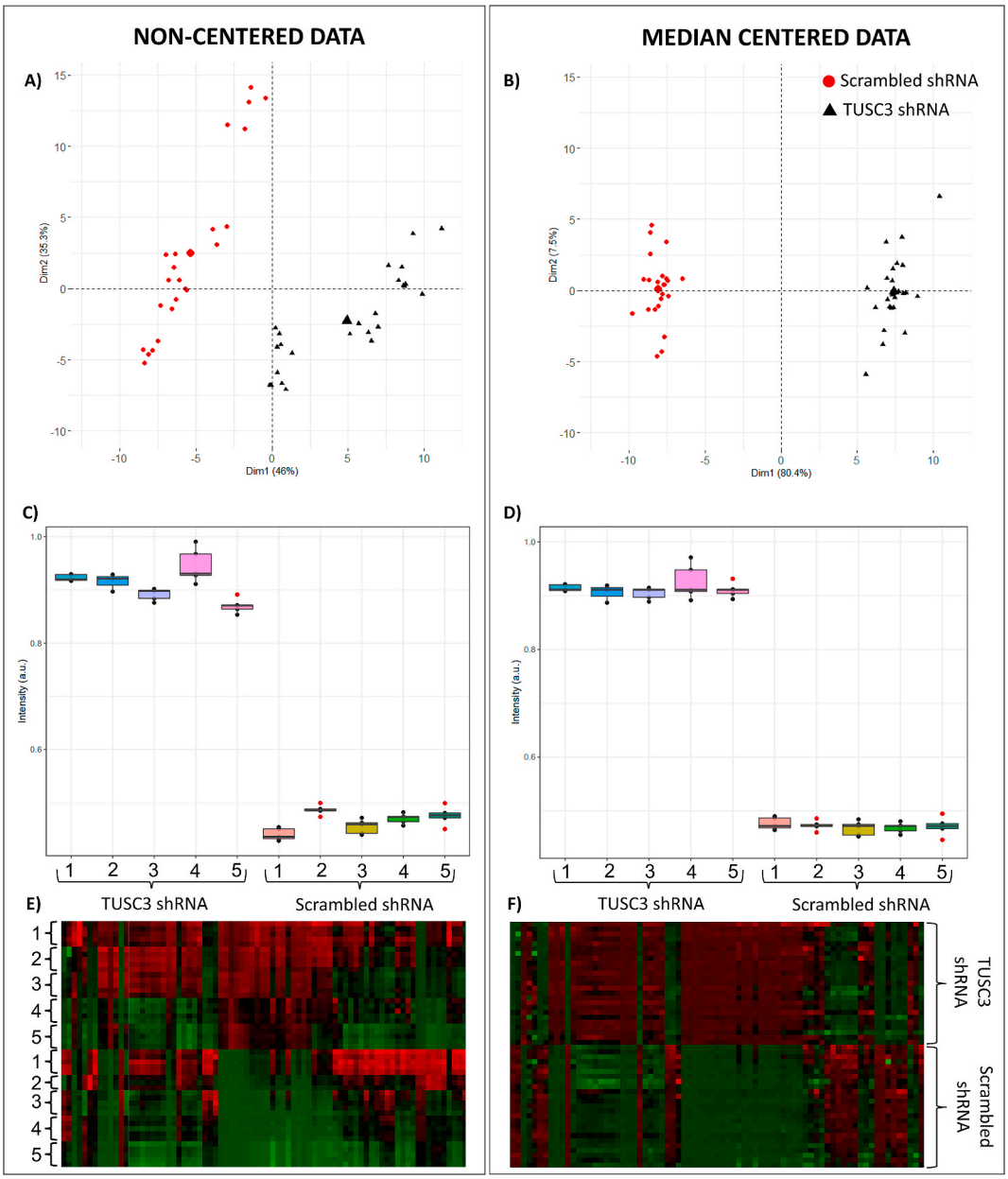


Fig. 4. Reduction of batch effects using the median-centered data; PCA of non-centered (**A**) and median-centered (**B**) spectral data recorded from $SKOV3_{scrambled\ shRNA}$ and $SKOV3_{TUSC3\ shRNA}$ in technical pentaplicates from five independent measurements. The intensity of peak at $m/z = 4939$ Da was used for demonstration of non-centered data (**C**) and median-centered data (**D**). Horizontal lines indicate median, boxes the first and third quartile, and whiskers the minimum and maximum values that fall within 1.5 times interquartile range. Heat maps of non-centered data (**E**) and median-centered data (**F**) were constructed from all peak intensities (feature matrix). Numbers 1–5 indicate replicates measured on different days.

clusters were generated from spectral data according to mixture compositions. The gradual joining of groups based on the increasing ratio of individual components in the population is shown. The associated heat map demonstrated a striking separation of data into seven distinct groups. Z-score normalized across groups was performed for heat map visualization.

To verify whether the ratio between $SKOV3_{scrambled\ shRNA}$ and $SKOV3_{TUSC3\ shRNA}$ can be predicted based on computationally processed spectral data, five machine learning (ML) algorithms – the Partial Least Square-Discrimination Analysis (PLS-DA), Decision Trees (DT), Random Forests (RF), Artificial neural networks (ANN) and k-Nearest Neighbors algorithm (k-NN), were investigated on the experimental spectral data. We used the peak intensities as the input variables for the ML algorithms. Architectures of classification algorithms were optimized to reduce the complexity of the classifier to improve prediction accuracy and hence to reduce the potential of overfitting. For DT, an algorithm with a root node, five internal nodes, seven leaf nodes, and three dominant variables (4909, 4939,

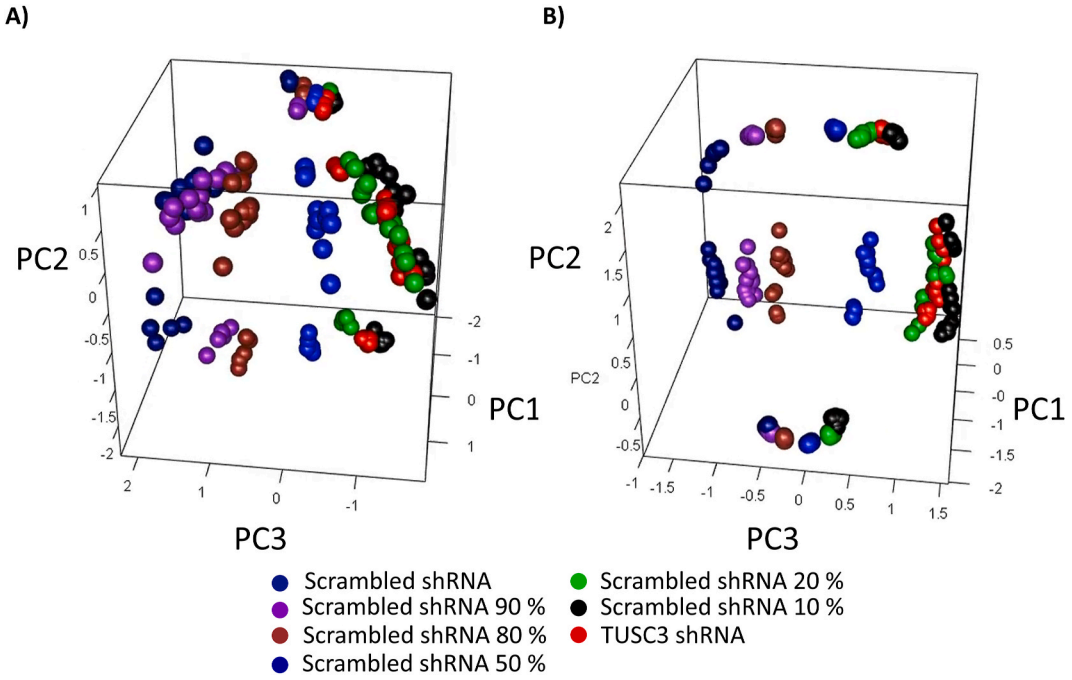


Fig. 5. Three-dimensional visualization of PCA of 175 mass spectra recorded from 7 binary cell mixtures for mean-centered data in PC1-PC3 space (**A**) and standard deviation-divided data (**B**).

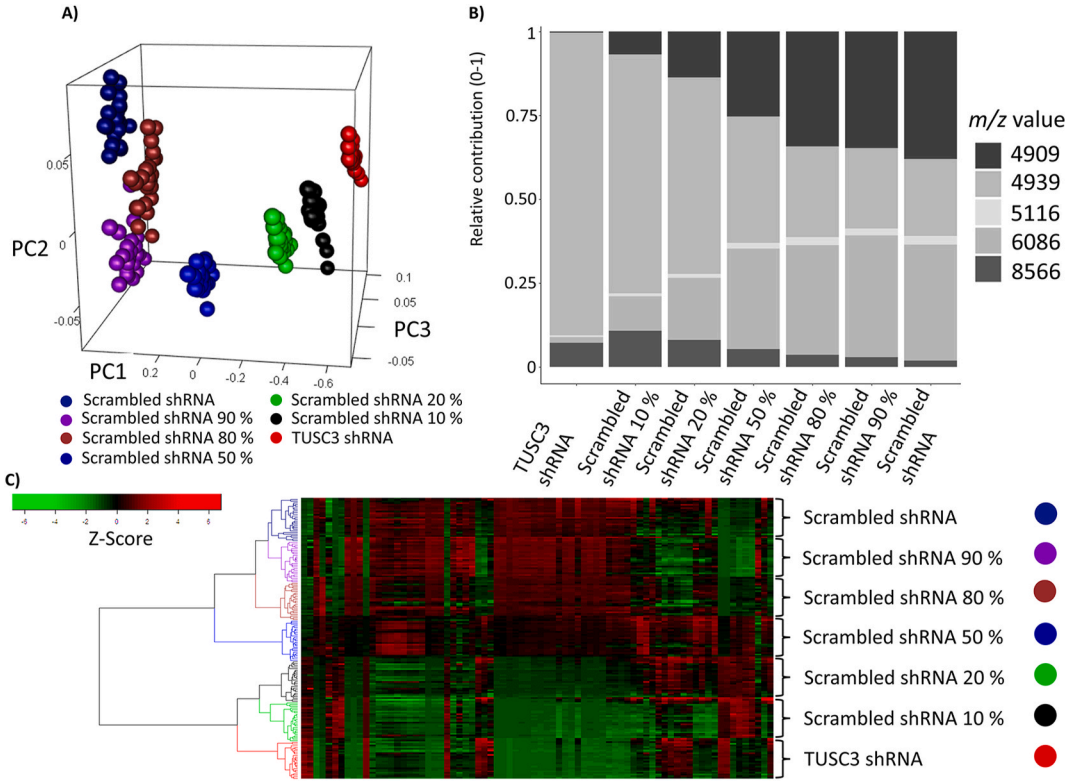


Fig. 6. Three-dimensional visualization of PCA of 175 mass spectra recorded from 7 binary cell population mixtures (**A**). Normalized stacked bar chart of selected *m/z* values as an average for individual cell mixtures (**B**). Heat map with hierarchical clustering analysis (Ward metric) of the 75 *m/z* signals (**C**).

and 8402) was used. For RF, 500 decision trees and a maximum of five variables for each tree were selected as the best structures. The variables 4909 and 8566 were identified as the most significant. For ANNs, the architecture containing five neurons in one hidden layer with 422 wt performed best. The training of ANN classifiers converged after 50 iterations. Performance for all optimized classifiers was compared and provided as the accuracy of prediction using 10 × repeated 5-fold CV (Fig. 7A). Results show that all classifiers had a significant performance in the dataset. For *k*-NN, 96.1 % accuracy was obtained (95 % confidence interval is used for all data, CI = 90.4–98.9 %). For DT 99.0 % accuracy was obtained (CI = 94.7–100.0 %). For RF 100.0 % accuracy was obtained (CI = 96.5–100.0 %). For ANN 100.0 % accuracy was obtained (CI = 96.5–100.0 %). Finally, PLS-DA with 100 % accuracy (CI = 96.5–100 %). The optimal number of components was determined based on the Q2Y parameter, which estimates the predictive performance of the model through the 5-fold cross-validation. The model's maximum Q2Y value indicates the number of components at which overfitting begins. A classifier based on PLS-DA was optimized for 15 components when the root mean square error (RMSE) reached a constant value and accuracy reached 100 % (Fig. 7B and **C**). In the case of pure populations only, the accuracy reached 100 % consistently (data not shown).

To avoid a systemic bias in the protocol or data analysis, we performed the in-lab validation. We prepared new samples from frozen cell pellets of the same biological batch, recorded 75 mass spectra in total, and used the trained classifiers to predict the composition of two-component mixtures. In the validation dataset, the accuracy of all classifiers was reduced. For the *k*-NN, 94.1 % accuracy was obtained (CI = 85.6–98.4 %). For DT 94.1 % accuracy was obtained (CI = 85.6–98.4 %) For the RF 98.5 % accuracy was obtained (CI = 92.1–100 %). For the ANN 98.5 % accuracy was obtained (CI = 94.7–100 %). Finally, PLS-DA with 100 % accuracy (CI = 94.7–100 %). When the trained PLS-DA classifier with optimized parameters was used to predict classification for the in-lab validation dataset, all data were correctly classified (75 from 7 cell mixtures) (Fig. 7D). The same result was achieved when ANN classifier was used (data not shown). Results for classification accuracy using PLS-DA classifiers with 2, 5, 8, and 10 components are provided as supplementary figure Fig. S5.

Next, to investigate, whether the classification algorithm is affected by dominant "marker" peak, we reduced the number of input variables by iterative random selection and performed the classification. Interestingly, even when the number of variables decreased from the original 75 variables to 20 only (10 random selections were performed), the accuracy of classification models did not decrease significantly. The accuracy of the PLS-DA model still reached 100 %. This finding suggests that the classification model is not influenced by a small number of dominant peaks and rather it is the cumulative effect of high number of variables with rather low weight that determines the robustness and accuracy of the model. It also indicates that the number of variables can be further reduced without compromising the performance of the algorithm. However, the reduction of variables entering the analysis can introduce a systemic bias that might be difficult to otherwise exclude. Similarly, when a high number of peak intensities (up to 300 were tested) was used for calculations, it resulted in comparable outcomes. The individual contribution of variables and the variable importance for the projection (VIP) were then calculated. VIP score then confirms that cumulative contribution of variables is necessary for the correct classification. The VIP plot is visualized in Fig. S6. Interestingly, when the mass spectra without median centering were used as inputs for ML classifiers, the performance of classifiers was not impaired significantly for pure $SKOV3_{scrambled\ shRNA}$ and $SKOV3_{TUSC3\ shRNA}$ cell samples. However, when mass spectra of the cell mixtures were used for classification without median centering, the classification error increased significantly. This suggests that ML applied on median-centered spectral data provides a highly robust tool for spectral analysis.

4. Conclusions

Our study reports the applicability of computationally processed intact cell MALDI-TOF MS spectral profiles for revealing changes in cell phenotypes induced by a single gene expression change. We investigated the machine learning algorithms (*k*-nearest neighbors, partial least squares discriminant analysis, decision tree, artificial neural network, and random forest) for the classification of two-component mixtures containing control and *TUSC3*-silenced cells. Our work also supports the use of mass window 2–10 kDa for robust and reproducible mass spectra as demonstrated in ovarian cancer cell line. We optimized preprocessing of mass spectra and construct classifiers that reveal even subtle changes in mass spectra related to corresponding to the targeted alteration of gene expression. In addition, analysis of peaks contributing to the spectral pattern can reveal new molecular biomarkers specific for cell types, metabolic states, and even genotypes.

Statements and declarations

The authors have no relevant financial or non-financial interests to disclose.

CRediT authorship contribution statement

Lukáš Pečinka: Writing – original draft, Methodology, Investigation. **Lukáš Moráň:** Validation, Methodology, Investigation. **Petra Kovačovicová:** Methodology, Investigation, Conceptualization. **Francesca Meloni:** Methodology, Investigation, Formal analysis. **Josef Havel:** Writing – review & editing, Validation, Funding acquisition, Formal analysis, Conceptualization. **Tiziana Pivetta:** Writing – review & editing, Validation, Formal analysis. **Petr Vaňhara:** Writing – review & editing, Writing – original draft, Supervision, Funding acquisition, Conceptualization.

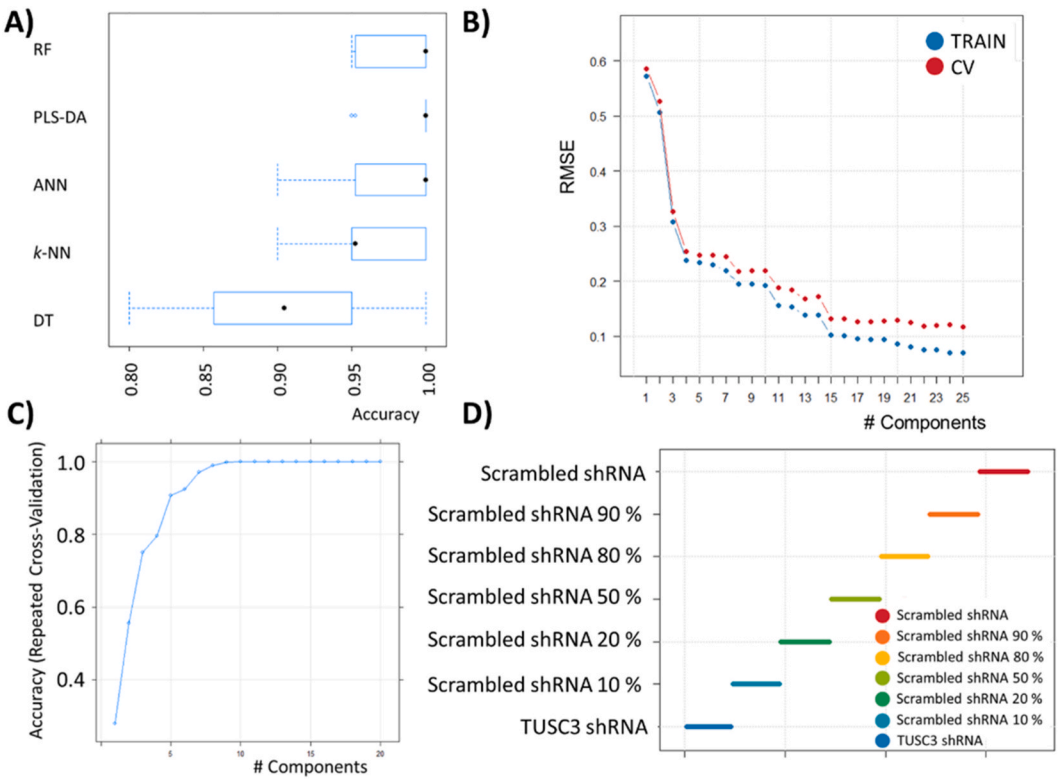


Fig. 7. Comparison of the prediction accuracy of the machine learning classifiers in binary cell mixtures (**A**). PLS-DA classifier Root Mean Square Error (RMSE) (**B**) and accuracy (**C**) visualized as functions of the number of input components. Plot documenting PLS-DA classification outputs for the validation dataset (**D**).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by Masaryk University (project nr. MUNI/A/1298/2022, MUNI/A/1301/2022 and MUNI/11/ACC/3/2022), and by the Czech Ministry of Health (NU23-08-00241).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e29936.

References

[1] M.E. Dueñas, E.A. Larson, Y.J. Lee, Toward mass spectrometry imaging in the metabolomics scale: increasing metabolic coverage through multiple on-tissue chemical modifications, Front. Plant Sci. 10 (2019) 1–11, https://doi.org/10.3389/fpls.2019.00860.
[2] C. Harkin, K.W. Smith, F.L. Cruickshank, C. Logan Mackay, B. Flinders, R.M.A. Heeren, T. Moore, S. Brockbank, D.F. Cobice, On-tissue chemical derivatization in mass spectrometry imaging, Mass Spectrom. Rev. 41 (2021) 662–694, https://doi.org/10.1002/mas.21680.
[3] M.Y. Ashfaq, D.A. Da'na, M.A. Al-Ghouti, Application of MALDI-TOF MS for identification of environmental bacteria: a review, J. Environ. Manag. 305 (2022) 114359–114370, https://doi.org/10.1016/j.jenvman.2021.114359.
[4] B. Munteanu, C. Hopf, Emergence of whole-cell MALDI-MS biotyping for high-throughput bioanalysis of mammalian cells? Bioanalysis 5 (2013) 885–893, https://doi.org/10.4155/bio.13.47.
[5] T.L. Williams, D. Andrzejewski, J.O. Lay, S.M. Musser, Experimental factors affecting the quality and reproducibility of MALDI TOF mass spectra obtained from whole bacteria cells, J. Am. Soc. Mass Spectrom. 14 (2003) 342–351, https://doi.org/10.1016/S1044-0305(03)00065-5.
[6] E. Valletta, L. Kučera, L. Prokeš, F. Amato, T. Pivetta, A. Hampl, J. Havel, P. Vaňhara, Multivariate calibration approach for quantitative determination of cell-line cross contamination by intact cell mass spectrometry and artificial neural networks, PLoS One 11 (2016) 1–14, https://doi.org/10.1371/journal.pone.0147414.

[7] P. Vaňhara, L. Kučera, L. Prokeš, L. Jurečková, E.M. Peña-Méndez, J. Havel, A. Hampl, Intact cell mass spectrometry as a quality control tool for revealing minute phenotypic changes of cultured human embryonic stem cells, Stem Cells Transl Med 7 (2018) 109–114, https://doi.org/10.1002/sctm.17-0107.
[8] A. Karger, B. Bettin, M. Lenk, T.C. Mettenleiter, Rapid characterisation of cell cultures by matrix-assisted laser desorption/ionisation mass spectrometric typing, J Virol Methods 164 (2010) 116–121, https://doi.org/10.1016/j.jviromet.2009.11.022.
[9] C.M. Buchanan, A.S. Malik, G.J.S. Cooper, Direct visualisation of peptide hormones in cultured pancreatic islet alpha- and beta-cells by intact-cell mass spectrometry, Rapid Commun. Mass Spectrom. 21 (2007) 3452–3458, https://doi.org/10.1002/rcm.3253.
[10] S.L. Kober, H. Meyer-Alert, D. Grienitz, H. Hollert, M. Frohme, Intact cell mass spectrometry as a rapid and specific tool for the differentiation of toxic effects in cell-based ecotoxicological test systems, Anal. Bioanal. Chem. 407 (2015) 7721–7731, https://doi.org/10.1007/s00216-015-8937-2.
[11] V.Z. Petukhova, A.N. Young, J. Wang, M. Wang, A. Ladanyi, R. Kothari, J.E. Burdette, L.M. Sanchez, Whole cell MALDI fingerprinting is a robust tool for differential profiling of two-component mammalian cell mixtures, J. Am. Soc. Mass Spectrom. 30 (2019) 344–354, https://doi.org/10.1007/s13361-018-2088-6.
[12] H. Kotasová, M. Capandová, V. Pelková, J. Dumková, Z. Koledová, J. Remšík, K. Souček, Z. Garlíková, V. Sedláková, A. Rabata, P. Vaňhara, L. Moráň, L. Pečinka, V. Porokh, M. Kučírek, L. Streit, J. Havel, A. Hampl, Expandable lung epithelium differentiated from human embryonic stem cells, Tissue Eng Regen Med 19 (2022) 1033–1050, https://doi.org/10.1007/s13770-022-00458-0.
[13] E. Valletta, L. Kučera, L. Prokeš, F. Amato, T. Pivetta, A. Hampl, J. Havel, P. Vaňhara, Multivariate calibration approach for quantitative determination of cell-line cross contamination by intact cell mass spectrometry and artificial neural networks, PLoS One 11 (2016), https://doi.org/10.1371/journal.pone.0147414.
[14] L. Soler, S. Uzbekova, E. Blesbois, X. Druart, V. Labas, Intact cell MALDI-TOF mass spectrometry, a promising proteomic profiling method in farm animal clinical and reproduction research, Theriogenology 150 (2020) 113–121, https://doi.org/10.1016/j.theriogenology.2020.02.037.
[15] E. Mohorko, R.L. Owen, G. Malojčić, M.S. Brozzo, M. Aebi, R. Glockshuber, Structural basis of substrate specificity of human oligosaccharyl transferase subunit N33/Tusc3 and its role in regulating protein N-glycosylation, Structure 22 (2014) 590–601, https://doi.org/10.1016/j.str.2014.02.013.
[16] K. Vaková, P. Horak, P. Vaňhara, TUSC3: functional duality of a cancer gene, Cell. Mol. Life Sci. 75 (2018) 849–857, https://doi.org/10.1007/s00018-017-2660-4.
[17] P. Vaňhara, P. Horak, D. Pils, M. Anees, M. Petz, W. Gregor, R. Zeillinger, M. Krainer, Loss of the oligosaccharyl transferase subunit TUSC3 promotes proliferation and migration of ovarian cancer cells, Int. J. Oncol. 42 (2013) 1383–1389, https://doi.org/10.3892/ijo.2013.1824.
[18] K. Kratochvílová, P. Horak, M. Ešner, K. Souček, D. Pils, M. Anees, E. Tomasich, F. Dráfi, V. Jurtíková, A. Hampl, M. Krainer, P. Vaňhara, Tumor suppressor candidate 3 (TUSC3) prevents the epithelial-to-mesenchymal transition and inhibits tumor growth by modulating the endoplasmic reticulum stress response in ovarian cancer cells, Int. J. Cancer 137 (2015) 1330–1340, https://doi.org/10.1002/ijc.29502.
[19] D. Pils, P. Horak, P. Vanhara, M. Anees, M. Petz, A. Alfanz, A. Gugerell, M. Wittinger, A. Gleiss, V. Auner, D. Tong, R. Zeillinger, E.I. Braicu, J. Sehouli, M. Krainer, Methylation status of TUSC3 is a prognostic factor in ovarian cancer, Cancer 119 (2013) 946–954, https://doi.org/10.1002/cncr.27850.
[20] S. Gibb, K. Strimmer, Maldiquant: A versatile R package for the analysis of mass spectrometry data, Bioinformatics 28 (2012) 2270–2271, https://doi.org/10.1093/bioinformatics/bts447.
[21] P.J. Rousseeuw, C. Croux, Alternatives to the median absolute deviation, J. Am. Stat. Assoc. 88 (1993) 1273, https://doi.org/10.2307/2291267.
[22] M.U.A. Bromba, H. Ziegler, Application hints for Savitzky-Golay digital smoothing filters, Anal. Chem. 53 (1981) 1583–1586, https://doi.org/10.1021/ac00234a011.
[23] C.G. Ryan, E. Clayton, W.L. Griffin, S.H. Sie, D.R. Cousens, SNIP, a statistics-sensitive background treatment for the quantitative analysis of PIXE spectra in geoscience applications, Nucl. Instrum. Methods Phys. Res. B. 34 (1988) 396–402, https://doi.org/10.1016/0168-583X(88)90063-8.
[24] N. Topić Popović, S.P. Kazazić, K. Bojanić, I. Strunjak-Perović, R. Čož-Rakovac, Sample preparation and culture condition effects on MALDI-TOF MS identification of bacteria: a review, Mass Spectrom. Rev. 42 (2023) 1589–1603, https://doi.org/10.1002/mas.21739.
[25] B. Balluff, C. Hopf, T. Porta Siegel, H.I. Grabsch, R.M.A. Heeren, Batch effects in MALDI mass spectrometry imaging, J. Am. Soc. Mass Spectrom. 32 (2021) 628–635, https://doi.org/10.1021/jasms.0c00393.
[26] E. Szájli, T. Fehér, K.F. Medzihradszky, Investigating the quantitative nature of MALDI-TOF MS, Mol. Cell. Proteomics 7 (2008) 2410–2418, https://doi.org/10.1074/mcp.M800108-MCP200.
[27] Y. Pan, S. Caudill, R. Li, K.L. Caldwell, Median and quantile tests under complex survey design using SAS and R, Comput. Methods Progr. Biomed. 176 (2017) 139–148, https://doi.org/10.1016/j.cmpb.2014.07.007.Median.