Data Article

# Quantitative proteomic dataset of the moss *Physcomitrium patens PSEP3* KO and OE mutant lines

Anna Mamaeva, Andrey Knyazev, Anna Glushkevich, Igor Fesenko*

*Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry of the Russian Academy of Sciences, Moscow, Russia*

## ARTICLE INFO

## ABSTRACT

Small open reading frames ($<$100 codons) that are located on long noncoding RNAs (lncRNAs) can encode functional microproteins. These microproteins are shown to play important roles in different cellular processes, such as cell proliferation, development and disease response [1–6]. However, there are only a few known lncRNA-encoded functional microproteins in plants. One such microprotein that was named PSEP3, was identified in the moss *Physcomitrium patens* by mass-spectrometry analysis. 57-aa PSEP3 contains Low Complexity Region (LCR) enriched with proline. We have previously shown that PSEP3 is translated in protonemata and gametophores of *P. patens*, and its knockout (KO line) or overexpression (OE line) affects protonemata growth [7]. We performed a quantitative proteomic analysis of the mutant lines with PSEP3 knockout and overexpression. 7-days old protonemata of wild type (WT line) and both mutant lines (KO and OE) were collected and used for iTRAQ-based proteomic experiments. LC-MS/MS data were processed using PEAKS Studio v.8 software with protein identification based on a Phytozome protein database. More analysis of PSEP3 effects on plant growth can be obtained in the paper published in Nucleic Acid Research [8].

---

* Corresponding author.
  *E-mail address:* moss@ibch.ru (I. Fesenko).
  *Social media:* 🐦 (I. Fesenko)

## Specifications Table

| | |
|---|---|
| Subject | Omics: Proteomics |
| Specific subject area | Plant quantitative proteomics |
| Type of data | Table |
| How data were acquired | Raw data were acquired with mass spectrometry using Q Exactive HF benchtop Orbitrap mass spectrometer (Thermo Fisher Scientific) and iTRAQ kit, analysis was performed using PEAKS Software 8.0 |
| Data format | Raw and analyzed data |
| Description of data collection | The protonemata of WT, PSEP3 KO and PSEP3 OE mutant lines were grown in 200 ml liquid BCD medium supplemented with 5 mM ammonium tartrate (BCDAT) during a 16 h photoperiod at 25 ℃ [9]. After 7 days of cultivation, protonemata were collected for analysis. The experiment was performed in three biological replicates. |
| Data source location | Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry of the Russian Academy of Sciences<br>Moscow<br>Russia |
| Data accessibility | Data identification number: 10.17632/NNCZRR9Y32.1 Anna Glushkevich, 2021. Quantitative proteomic dataset of the moss Physcomitrium patens PSEP3 KO and OE mutant lines.<br>Direct URL to data:<br>https://data.mendeley.com/datasets/nnczrr9y32/1 |
| Related research article | I. Fesenko, S.A. Shabalina, A. Mamaeva, A. Knyazev, A. Glushkevich, I. Lyapina, R. Ziganshin, S. Kovalchuk, D. Kharlampieva, V. Lazarev, M. Taliansky, E.V. Koonin, A vast pool of lineage-specific microproteins encoded by long non-coding RNAs in plants, Nucleic Acids Research, 49, (2021) 10328–10346, 10.1093/nar/gkab816 |

## Value of the Data

- These data are useful for comparative studies of cellular pathways regulated by lncRNA-encoded microproteins in plants and animals. The list of differentially regulated proteins identified in *PSEP3* mutant lines can be compared with microprotein-regulated genes from other organisms to shed light on mechanisms of the appearance and evolution of microproteins.
- These data could be useful for large-scale studies of side effects of CRISPR-Cas9-derived mutations in plants.
- The presented data can be re-analyzed with databases containing previously unannotated protein sequences such as other lncRNA-encoded microproteins, transposons, and alternative open reading frames. It will help to understand the role of microproteins in cellular processes better.

## 1. Data Description

Here we present a quantitative proteomics dataset of *PSEP3* KO and OE mutant lines. For the proteomic experiments, moss protonemata were grown at a liquid BCD-AT medium for 7 days (Fig. 1). We used isobaric tags for relative and absolute quantitation (iTRAQ) for accurate detection of the quantitative changes at the protein level in mutant lines. The raw LC-MS/MS data is available in the Mendeley Data database (10.17632/NNCZRR9Y32.1). The protein identification and quantification were conducted using PEAKS Studio 8.0 software.
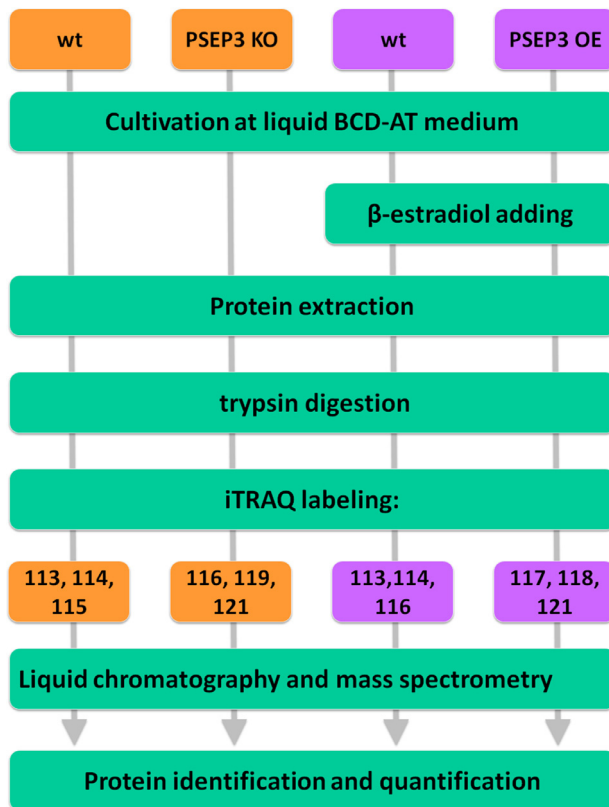
**Fig. 1.** Overview of the experimental workflow.

**Table 1**
Characteristics of the proteomic datasets.

|  | PSEP3 KO dataset | PSEP3 OE dataset |
|---|---|---|
| Number of MS/MS scans | 45815 | 54029 |
| Peptide-Spectrum Matches | 17784 | 27862 |
| Peptide sequences | 11170 | 16724 |
| Protein groups | 2710 | 2873 |
| Proteins | 3461 | 3511 |

In total, we identified 2710 protein groups supported by 17748 peptide-spectrum matches (PSMs) in *PSEP3* KO mutant plants (Table 1). 56 protein groups were differentially expressed (FC > 1.2, *P* < 0.01), 21 of these were upregulated in comparison to wild type plants (Tables are available on Mendeley Data, 10.17632/NNCZRR9Y32.1). The distributions of peptide lengths, coverage, protein mass and peptide number in analysed samples are shown at Fig. 2. According to our previous study [8], the overexpression of PSEP3 induces cell death in 48 h. Therefore, we studied the effect of short-term PSEP3 overexpression in mutant lines using the $\beta$-estradiol induction system [10]. We analyzed proteomic changes in 4 h after *PSEP3* induction with $\beta$-estradiol. In total, we observed 2873 protein groups supported by 27862 PSMs in this sample (Table 1). We identified 167 differentially expressed protein groups (DEPs) after induction of PSEP3 overexpression (Tables are available on Mendeley Data, 10.17632/NNCZRR9Y32.1).
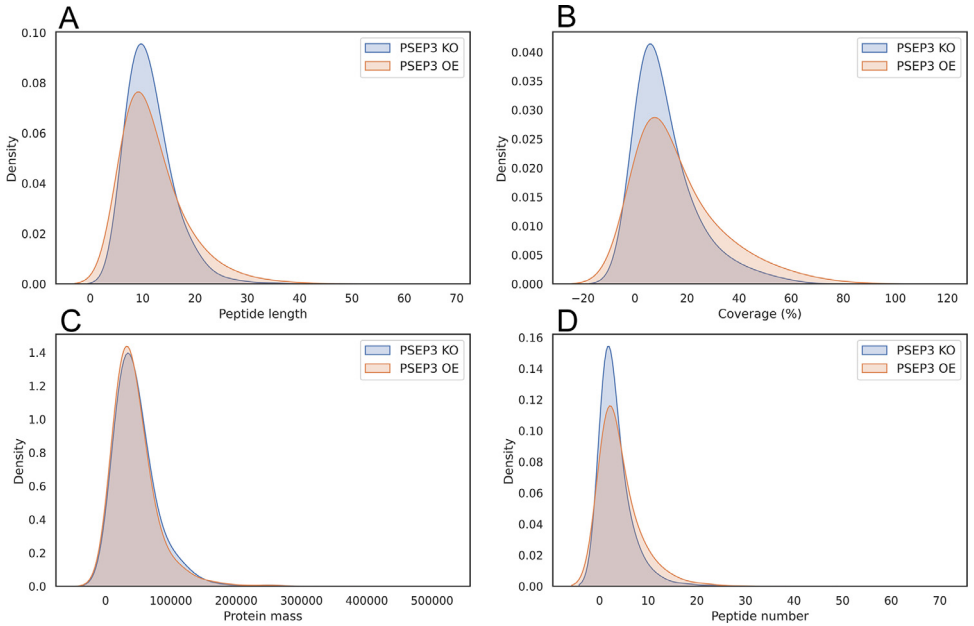
**Fig. 2.** Quality control metrics of the proteomic datasets. (**A**) The length distributions of peptides; (**B**) Protein sequence coverage distributions; (**C**) Protein mass distributions and (**D**) Distributions of the unique peptide numbers.
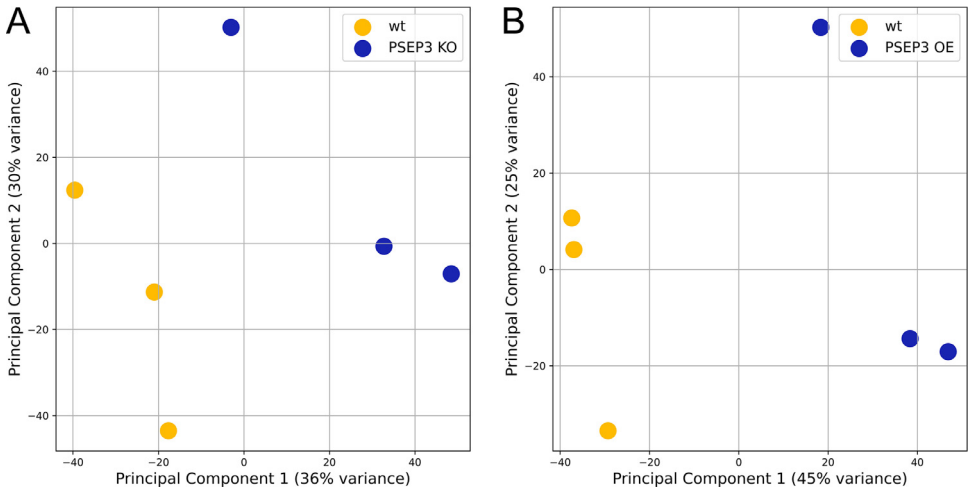


**Fig. 3.** Principal component analysis (PCA) of wild type with PSEP3 KO (**A**) and PSEP3 OE (**B**) mutant samples. Wild type shown by orange and mutants shown by blue. The PCA analysis included all protein groups and was performed in the Python library sklearn.

Principal component analysis (PCA) based on intensities of all identified protein groups is shown in Fig. 3.

We found that some photosynthetic proteins were downregulated in *PSEP3* KO plants but up-regulated in PSEP3 OE plants (Tables are available on Mendeley Data, 10.17632/NNCZRR9Y32.1). Such DEPs as catalase, xyloglucan endo-transglycosylase and metacaspase-4-related protein downregulated in *PSEP3* KO plants, whereas dynamin-related protein 1C was upregulated in

*PSEP3* OE plants. These proteins are involved in antioxidant defense, cell wall structure, cell death and organelle functioning [11–13].

## 2. Experimental Design, Materials and Methods

### 2.1. Plant material

The moss *Physcomitrium patens* subsp. *patens* ("Gransden 2004", Freiburg) of wild type and *PSEP3* KO and OE mutant lines produced earlier using CRISPR/Cas9 technology [7] were used in this study. The moss protonemata were grown in 200 ml liquid BCD medium supplemented with 5 mM ammonium tartrate (BCD-AT) during a 16 h photoperiod at 25 °C for 7 days [9].

### 2.2. Protein extraction and trypsin digestion

Proteins were extracted with the phenol extraction method [14]. Samples were homogenized in 3 ml ice-cold extraction buffer (500 mM Tris–HCl, pH 8.0, 50 mM EDTA, 700 mM sucrose, 100 mM KCl, 1 mM PMSF, 1 mM DTT), followed by 10 min incubation on ice. The 3 ml of ice-cold Tris–HCl (pH 8.0)-saturated phenol was added, and the mixture was vortexed and incubated for 10 min with shaking. After centrifugation (10 min, 5500 × g, 4 °C), the phenol phase was collected and re-extracted with a 3 ml of extraction buffer. Proteins were precipitated with 8 ml of ice-cold 0.1 M ammonium acetate in methanol overnight at −20 °C. The samples were centrifuged (10 min, 5500 × g, 4 °C) and pellets were rinsed with centrifugation (10 min, 5500 × g, 4 °C) by ice-cold 0.1 M ammonium acetate in methanol three times and with ice-cold acetone once. The resulting pellets were dried and dissolved in 8 M urea, 2 M thiourea and 10 mM Tris. Proteins were quantified by Bradford protein assay (Bio-Rad, Hercules, CA USA). The 100 μg of proteins were reduced by 5 mM DTT for 30 min at 50 °C and alkylated by 10 mM iodoacetamide for 20 min at room temperature. Proteins were dissolved in 40 mM ammonium bicarbonate and digested by 1 μg sequence-grade modified trypsin (Promega, Madison, WI, USA) at 37 °C overnight. The reaction was stopped by adding trifluoroacetic acid (TFA) to the final concentration of 1%. 20 μg of each sample was desalted by Empore octadecyl C18 extraction disks (Supelco, USA) and then was dried in a vacuum concentrator. iTRAQ labeling (Applied Biosystems, Foster City, CA, USA) was conducted according to the manufacturer's manual. Proteins were labeled with the iTRAQ tags as follows: wild type - 113-115, *PSEP3* KO - 116,119,121 ones; wild type (with estradiol) - 113, 114, 116, OE PEP3 (with estradiol) - 117,118, 121 ones. Samples were mixed, vacuum dried and dissolved in 1% TFA. The mixture was desalted on SCX extraction disks (Supelco, USA): columns were washed by 0.1% TFA and eluted by 5% ammonium hydroxide in 80% acetonitrile. Each sample was dried in a vacuum concentrator and dissolved in 3% acetonitrile with 0.1% trifluoroacetic acid.

### 2.3. Liquid chromatography and mass spectrometry

LC-MS/MS analysis was conducted as described in our previous studies [15,16]. Peptides were separated on Acclaim PepMap 100 C18 (75 μm × 50 cm) (Thermo Fisher Scientific). Reverse-phase chromatography was performed with an Ultimate 3000 Nano LC System (Thermo Fisher Scientific), which was coupled to the Q Exactive HF benchtop Orbitrap mass spectrometer (Thermo Fisher Scientific) via a nanoelectrospray source (Thermo Fisher Scientific). Peptides in 5 μL of loading buffer (3% (vol/vol) acetonitrile, 0.1% (vol/vol) TFA in Milli-Q deionized water) were loaded on a trapping column PepMap 100 C18 (0.1 × 20 mm) (Thermo Fisher Scientific) at a flow rate of 5 μL/min for 6 min. NanoLC pump mobile phases were: A – (2% (vol/vol) acetonitrile, 0.1% (vol/vol) formic acid in Milli-Q deionized water; B - (80% (vol/vol) acetonitrile,

0.1% (vol/vol) formic acid, 19.9% (vol/vol) milli-Q deionized water. Peptides were eluted from the trapping column with a linear gradient: 5–28% B for 90 min; 28-45% B for 20 min, and 45–100% B for 7 min at a flow rate of 350 nL/min. After each gradient, the column was washed with 100% buffer B for 5 min and reequilibrated with buffer A for 10 min. Peptides were analyzed on a mass spectrometer, with one full scan (375–1400 m/z, R = 120,000 at 200 m/z) at a target of $3{*}10^6$ ions and max ion fill time 50 ms, followed by up to 15 data-dependent MS/MS scans with higher-energy collisional dissociation (HCD) (target $1{*}10^5$ ions, max ion fill time 100 ms, isolation window 1.2 m/z, normalized collision energy (NCE) 32%), detected in the Orbitrap (R = 30,000 at fixed first mass 100 m/z). Other settings: charge exclusion – unassigned, 1, > 6; peptide match – preferred; exclude isotopes – on; dynamic exclusion – 60 s was enabled.

### 2.4. Protein identification and quantification

Tandem mass spectra were analysed by PEAKS Studio version 8.0 software (Bioinformatics Solutions Inc., Waterloo, Canada) [17]. The custom database was built from the Phytozome database *P. patens* combined with chloroplast and mitochondrial proteins (33,053 records) (Supplementary file 1). The database search was performed with the following parameters: a fragmentation mass tolerance of 0.05 Da; parent ion tolerance of 10 ppm; fixed modification – carbamidomethylation; variable modifications – oxidation (M) and acetylation (Protein N-term). The resulting protein list was filtered by a 1% false discovery rate (FDR).

PEAKS Q was used for iTRAQ quantification. Normalization was performed by averaging the abundance of all peptides. Median values were used for averaging. Given that iTRAQ quantification typically underestimates the degree of real fold changes between two samples, differential protein screening was performed using a fold change ratio $\geqslant$ 1.20 (for upregulated DEPs) or $\leqslant$ 0.83 (for downregulated DEPs) and Significance $\geqslant$ 15. Protein significance analysis was performed in PEAKS 8.0 (Bioinformatics Solutions Inc., Waterloo, Canada) [17]. A two-tailed t-test was used for calculation of significant differences in protein ratios ($P < 0.05$).

### 2.5. Statistical analysis

Two-component principal component analysis (PCA) was performed based on standardized intensity iTRAQ values of each protein group in the samples using the scikit-learn Python package [18]. Standardization and decomposition were performed in the scikit-learn package with default parameters. The columns "Coverage (%)", "Avg. Mass" and "#Peptides" from PEAKs result tables were used for analysis of quality control metrics of the proteomic datasets. Peptide lengths were counted as differences between Start and Stop peptide positions. Visualizations were made in Python using module seaborn 0.11.1 [19].

### Ethics Statement

Not applicable.

### Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have, or could be perceived to have, influenced the work reported in this article.

## CRediT Author Statement

**Anna Mamaeva:** Investigation, Writing – original draft; **Andrey Knyazev:** Investigation; **Anna Glushkevich:** Writing – review & editing, Visualization; **Igor Fesenko:** Formal analysis, Writing – review & editing, Data curation.

## Acknowledgments

## Supplementary Materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.dib.2021.107715.

## References

[1] M. Polycarpou-Schwarz, M. Gross, P. Mestdagh, J. Schott, S.E. Grund, C. Hildenbrand, J. Rom, S. Aulmann, H.P. Sinn, J. Vandesompele, S. Diederichs, The cancer-associated microprotein CASIMO1 controls cell proliferation and interacts with squalene epoxidase modulating lipid droplet formation, Oncogene 37 (2018) 4750–4768, doi:10.1038/s41388-018-0281-5.

[2] B. Guo, S. Wu, X. Zhu, L. Zhang, J. Deng, F. Li, Y. Wang, S. Zhang, R. Wu, J. Lu, Y. Zhou, Micropeptide CIP2A-BP encoded by LINC00665 inhibits triple-negative breast cancer progression, EMBO J. 39 (2020) e102190, doi:10.15252/embj.2019102190.

[3] J.Z. Huang, M. Chen, D. Chen, X.C. Gao, S. Zhu, H.Y. Huang, M. Hu, H.F. Zhu, G.R. Yan, A peptide encoded by a putative lncRNA HOXB-AS3 suppresses colon cancer growth, Mol. Cell 68 (2017) 171, doi:10.1016/j.molcel.2017.09.015.

[4] M. Rubtsova, Y. Naraykina, D. Vasilkova, M. Meerson, M. Zvereva, V. Prassolov, V. Lazarev, V. Manuvera, S. Kovalchuk, N. Anikanov, et al., Protein encoded in human telomerase RNA is involved in cell protective pathways, Nucleic Acids Res. 46 (2018) 8966–8977, doi:10.1093/nar/gky705.

[5] A. Chugunova, E. Loseva, P. Mazin, A. Mitina, T. Navalayeu, D. Bilan, P. Vishnyakova, M. Marey, A. Golovina, M. Serebryakova, P. Pletnev, M. Rubtsova, W. Mair, A. Vanyushkina, P. Khaitovich, V. Belousov, M. Vysokikh, P. Sergiev, O. Dontsova, LINC00116 codes for a mitochondrial peptide linking respiration and lipid metabolism, Proc. Natl. Acad. Sci. U. S. A. 116 (2019) 4940–4945, doi:10.1073/pnas.1809105116.

[6] A. Matsumoto, A. Pasut, M. Matsumoto, R. Yamashita, J. Fung, E. Monteleone, A. Saghatelian, K.I. Nakayama, J.G. Clohessy, P.P. Pandolfi, mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide, Nature 541 (2017) 228–232, doi:10.1038/nature21034.

[7] I. Fesenko, I. Kirov, A. Kniazev, R. Khazigaleeva, V. Lazarev, E. Grafskaia, V. Zgoda, I. Butenko, G. Arapidi, A. Mamaeva, V. Ivanov, V. Govorun, Distinct types of short open reading frames are translated in plant cells, Genome Res. 29 (2019) 1464–1477, doi:10.1101/gr.253302.119.

[8] I. Fesenko, S.A. Shabalina, A. Mamaeva, A. Knyazev, A. Glushkevich, I. Lyapina, R. Ziganshin, S. Kovalchuk, D. Kharlampieva, V. Lazarev, M. Taliansky, E.V. Koonin, A vast pool of lineage-specific microproteins encoded by long noncoding RNAs in plants, Nucleic Acids Res. (2021), doi:10.1093/nar/gkab816.

[9] D.J. Cove, P.F. Perroud, A.J. Charron, S.F. McDaniel, A. Khandelwal, R.S. Quatrano, Culturing the moss Physcomitrella patens, Cold Spring Harb. Protoc. 2009 (2009) db.prot5136.

[10] M. Kubo, A. Imai, T. Nishiyama, M. Ishikawa, Y. Sato, T. Kurata, Y. Hiwatashi, R. Reski, M. Hasebe, System for Stable β-Estradiol-Inducible Gene Expression in the Moss Physcomitrella patens, PLoS One 8 (2013) e77356, doi:10.1371/journal.pone.0077356.

[11] L. Tsiatsiani, F. Van Breusegem, P. Gallois, A. Zavialov, E. Lam, P.V. Bozhkov, Metacaspases, Cell Death Differ. 18 (2011) 1279–1288, doi:10.1038/cdd.2011.66.

[12] R. Jilly, N.Z. Khan, H. Aronsson, D. Schneider, Dynamin-like proteins are potentially involved in membrane dynamics within chloroplasts and cyanobacteria, Front. Plant Sci. 9 (2018) 206, doi:10.3389/fpls.2018.00206.

[13] S. Miyagishima, K. Nishida, T. Mori, M. Matsuzaki, T. Higashiyama, H. Kuroiwa, T. Kuroiwa, A plant-specific dynamin-related protein forms a ring at the chloroplast division site, Plant Cell 15 (2003) 655–665, doi:10.1105/tpc.009373.

[14] M. Faurobert, C. Mihr, N. Bertin, T. Pawlowski, L. Negroni, N. Sommerer, M. Causse, Major proteome variations associated with cherry tomato pericarp development and ripening, Plant Physiol. 143 (2007) 1327–1346.

[15] I. Fesenko, N. Spechenkova, A. Mamaeva, A.V. Makhotenko, A.J. Love, N.O. Kalinina, M. Taliansky, Role of the methionine cycle in the temperature-sensitive responses of potato plants to potato virus Y, Mol. Plant Pathol. 22 (2021) 77–91.

[16] N. Spechenkova, I.A. Fesenko, A. Mamaeva, T.P. Suprunova, N.O. Kalinina, A.J. Love, M. Taliansky, The resistance responses of potato plants to potato virus Y are associated with an increased cellular methionine content and an altered SAM:SAH methylation index, Viruses 13 (2021) 955, doi:10.3390/v13060955.

[17] B. Ma, K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby, G. Lajoie, PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry, Rapid Commun. Mass Spectrom. 17 (2003) 2337–2342.

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., Scikit-learn: machine learning in python, J. Mach. Learn. Res. 12 (Oct) (2011) 2825–2830.

[19] M. Waskom, seaborn: statistical data visualization, J. Open Source Softw. 6 (2021) 3021.