

# LORA, Lipid Over-Representation Analysis Based on Structural Information

Michaela Vondrackova, Dominik Kopczynski, Nils Hoffmann, and Ondrej Kuda\*



Cite This: *Anal. Chem.* 2023, 95, 12600–12604



Read Online

ACCESS |



Metrics & More

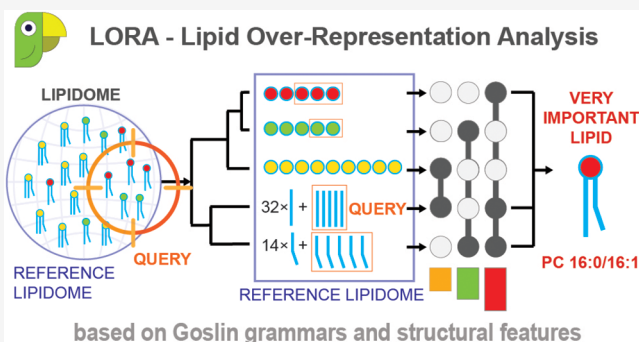


Article Recommendations



Supporting Information

**ABSTRACT:** With the increasing number of lipidomic studies, there is a need for an efficient and automated analysis of lipidomic data. One of the challenges faced by most existing approaches to lipidomic data analysis is lipid nomenclature. The systematic nomenclature of lipids contains all available information about the molecule, including its hierarchical representation, which can be used for statistical evaluation. The Lipid Over-Representation Analysis (LORA) web application (<https://lora.metabolomics.fgu.cas.cz>) analyzes this information using the Java-based Goslin framework, which translates lipid names into a standardized nomenclature. Goslin provides the level of lipid hierarchy, including information on headgroups, acyl chains, and their modifications, up to the “complete structure” level. LORA allows the user to upload the experimental query and reference data sets, select a grammar for lipid name normalization, and then process the data. The user can then interactively explore the results and perform lipid over-representation analysis based on selected criteria. The results are graphically visualized according to the lipidome hierarchy. The lipids present in the most over-represented terms (lipids with the highest number of enriched shared structural features) are defined as Very Important Lipids (VILs). For example, the main result of a demo data set is the information that the query is significantly enriched with “glycerophospholipids” containing “acyl 20:4” at the “sn-2 position”. These terms define a set of VILs (e.g., PC 18:2/20:4;O and PE 16:0/20:4(5,8,10,14);OH). All results, graphs, and visualizations are summarized in a report. LORA is a tool focused on the smart mining of lipidomics data sets to facilitate their interpretation at the molecular level.



## INTRODUCTION

Recent advances in analytical techniques and their routine use in screening pipelines push forward our ability to provide a full structural characterization of lipid species in complex biological matrices. The stereospecifically numbered (*sn*) position of the acyl/alkyl chain on the glycerol backbone,<sup>1</sup> double bond position and stereochemistry within acyl/alkyl chains,<sup>2,3</sup> and configuration of chiral centers<sup>4</sup> can be assigned using a combination of advanced separation and ion activation techniques.<sup>5</sup> The emerging challenge is to correctly and systematically annotate the lipid species<sup>6</sup> and consider the high structural diversity of modified lipid species, generally referred to as the “epilipidome”.<sup>7</sup> Correctly annotated and standardized lipid data sets represent an information source for further FAIR data mining.

Over-Representation Analysis (ORA) is a simple statistical method that determines whether an a priori-defined set of variables is more present (over-represented) in a subset of variables than would be expected by chance. Two main bioinformatics tools for over-representation analysis of lipidomics data sets are available: (1) LION,<sup>8</sup> a lipid ontology tool that associates >50,000 lipid species to biophysical, chemical, and cell biological features and (2) Lipid Mini-On,

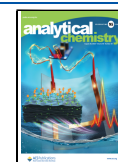
an open-source tool that performs lipid enrichment analyses and visualizations of lipidomics data. Both tools use custom-defined lipid databases, specific nomenclatures, and parsing functions to mine data from lipid names. However, these tools lack the power to exploit the hierarchical nature of lipid structures (e.g., *sn* position, double bond position) and leave the data unmined.

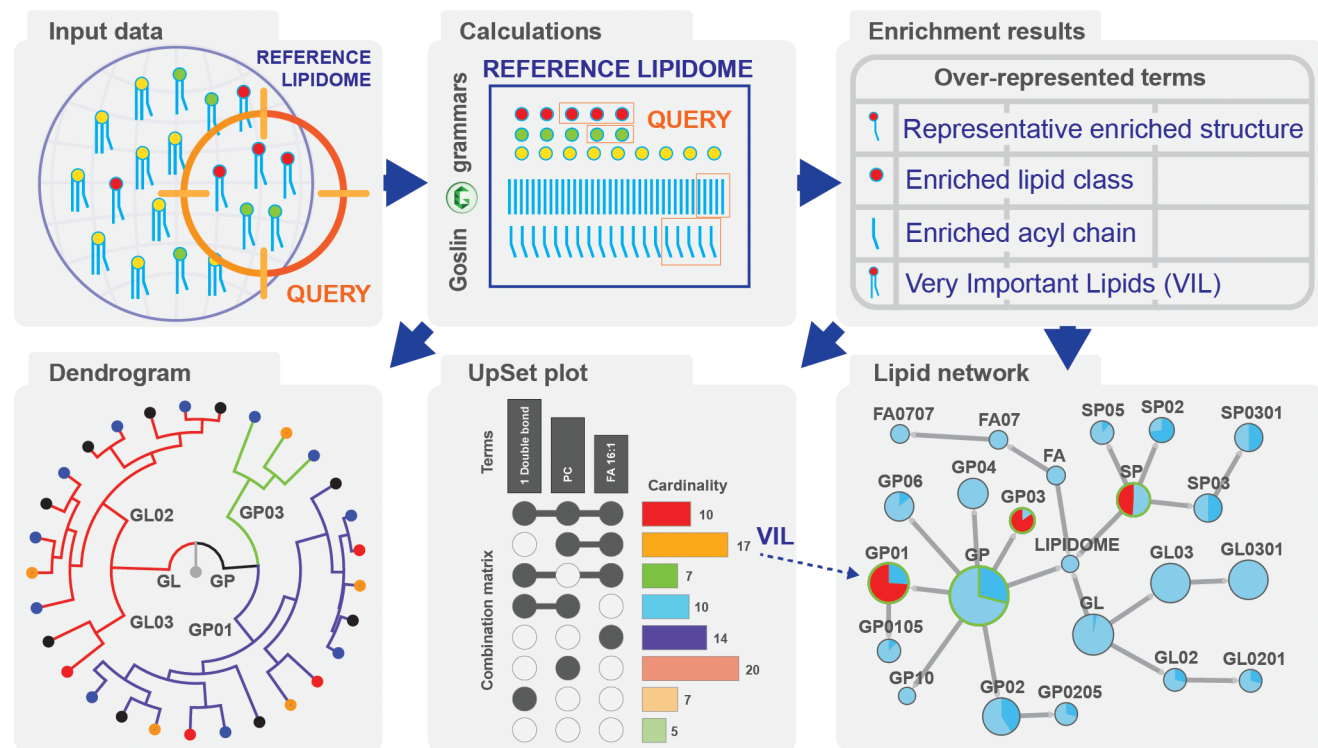
The goal of the LORA project was to build a bioinformatics tool based on Goslin, a systematic grammar-based lipid library, to facilitate the statistical evaluation of lipid structural information and to support international standardization of lipidomics nomenclature.<sup>6,10,11</sup>

**Received:** May 11, 2023

**Accepted:** August 1, 2023

**Published:** August 16, 2023





**Figure 1.** LORA pipeline. User data (reference and query lipid names) are processed using Goslin. The lipidome is visualized as a dendrogram, and LORA is performed. Enrichment results are summarized in a table, which is further processed to a lipid network and an UpSet plot.

## METHODS

**Lipid Identifiers and Goslin.** Bioinformatic tools for lipid ORA require lipid names or database identifiers. We used jGoslin, the Java implementation of Goslin,<sup>11</sup> which parses the submitted lipid names and translates them into a normalized hierarchical representation (Table S1). jGoslin supports lipid names based on LIPID MAPS, SwissLipids, HMDB, and Shorthand nomenclature. Lipid classes implemented in Goslin are periodically updated as new lipid classes are discovered.<sup>11</sup> Epilipidome nomenclature can be converted to a compatible format by LipidLynxX.<sup>12</sup>

**Data Sets.** The lipid data sets used in this work refer to four published data sets from lipidomic studies in humans: (1) DEMO 1: Cachexia, body weight stable vs cachectic patients (human epicardial adipose tissue),<sup>13</sup> (2) DEMO 2: AdipoAtlas, obese vs lean patients (human white adipose tissue),<sup>14</sup> (3) DEMO 3: Lipid Mini-On demo data set (human lung endothelial lipidome),<sup>9</sup> (4) technical DEMO 4: oxidized membrane lipids (human platelets) at “Complete structure level”, combined with phospholipids at “Structure defined” level (mouse liver), and a Goslin performance test file.<sup>3,10,15</sup> A list of query lipids, the whole reference lipidome (codomain), LORA manual, and the LORA report for each data set is available in the LORA web application (<https://lora.metabolomics.fgu.cas.cz>). The LORA manual describes how to prepare query and reference lipidome lists.

**Implementation of ORA.** Two enrichment tests were used to perform ORA: (1) Fisher exact test (`scipy.stats.fisher_exact`) with “two-sided”, “less”, “greater” alternatives and (2) hypergeometric test (`scipy.stats.hypergeom`) from SciPy.<sup>16</sup> Multiple comparisons were adjusted using the Bonferroni, Holm–Bonferroni, or Benjamini–Hochberg procedure (the classic False Discovery Rate, FDR).<sup>17,18</sup> The ORA was

performed at each level of the nomenclature hierarchy and on structural features provided by Goslin to correct for the set size effect.<sup>19</sup> Additional parameters (grouped number of carbon atoms per acyl chain: less than 16, 16–18, and more than 18; and the number of double bonds: 0, saturated; 1, monounsaturated; 2 and more, polyunsaturated) were included to facilitate biological interpretation of the data. UpSet plots provide visualization of term intersections.<sup>20</sup> The visualization of nomenclature levels was implemented via Biopython.<sup>21</sup> The lipidome was converted to phyloXML format, an XML language designed to describe phylogenetic trees (or networks) and associated data,<sup>22</sup> and the pseudophylogenetic distances (radii of the spheres/levels) were optimized for the hierarchical structure of the lipidome. Only the CATEGORY and CLASS levels are labeled to optimize space use and prevent overlapping labels. The Cytoscape network was built to visualize the quantitative data and statistics.<sup>23</sup>

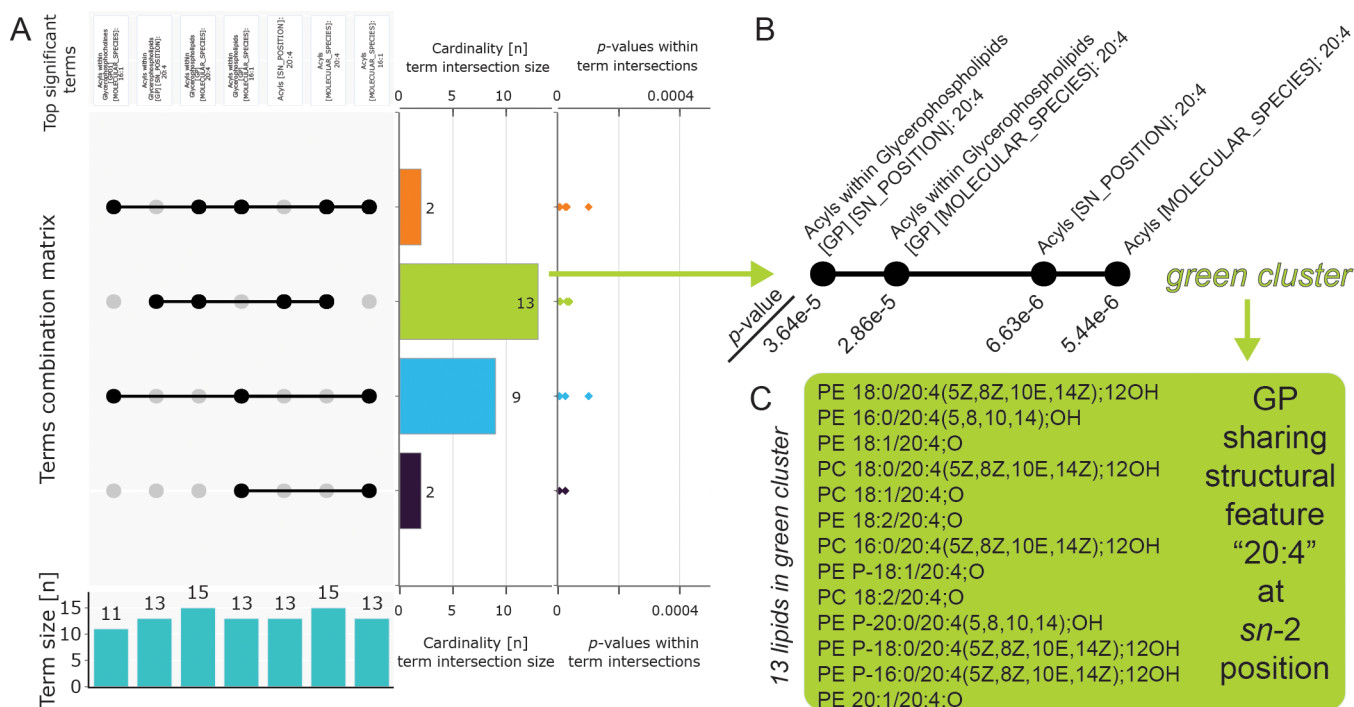
## RESULTS

**LORA, a Web-Based Tool for Lipid Over-Representation Analysis.** We developed LORA as a web-based interactive tool using Python 3.10, Dash, and Plotly packages.<sup>24,25</sup> The application uses a tabular layout. The query and reference lists of lipids are uploaded, processed, and parsed via Goslin. Users can change the ORA parameters, perform the analysis, interact with the plots, and download the report (Figure 1). LORA is available as a web service (<https://lora.metabolomics.fgu.cas.cz>), and the source code is hosted at GitHub <https://github.com/IPHYS-Bioinformatics/LORA> under the terms of liberal open source licenses.

The primary output of ORA is a set of over-represented terms. The terms are defined based on the nomenclature levels and structural characteristics of the lipidome parsed by Goslin,

Table 1. Summary of Over-Represented Terms from DEMO 4 Calculated at Alpha Level 0.001 with FDR Correction

Term (Group/Classifier)	Goslin Level	No. Query	No. Reference	p-value	Odds Ratio	FDR
Acyls 16:1	MOLECULAR SPECIES	13/68	18/556	4.30e-06	7.0646	0
Acyls 20:4	MOLECULAR SPECIES	15/68	26/556	5.44e-06	5.7692	0
Acyls 20:4	SN POSITION	13/68	19/556	6.63e-06	6.6804	0.0001
Acyls 16:1 within Glycerophospholipids [GP]	MOLECULAR SPECIES	13/68	17/448	2.36e-05	5.9925	0.0002
Acyls 20:4 within Glycerophospholipids [GP]	MOLECULAR SPECIES	15/68	24/448	2.86e-05	5.0000	0.0002
Acyls 20:4 within Glycerophospholipids [GP]	SN POSITION	13/68	18/448	3.64e-05	5.6465	0.0004
Acyls 16:1 within Glycerophosphocholines [GP01]	MOLECULAR SPECIES	11/26	14/148	1.00e-4	7.0190	0.0005



**Figure 2.** Interpretation of the LORA results. (A) UpSet plot showing the intersection of over-represented terms (structural characteristics) from Table 1. Cardinality is sorted according to the total number of term intersections. (B) Intersection of four terms defining the green cluster. (C) Green cluster containing 13 lipids which share a structural feature "20:4" at the *sn*-2 position.

e.g., "Total number of carbon atoms", "Fatty acid #1 *sn* position", Total #O, etc.

LORA is the first tool that uses smart text mining and extracts all available structural information from a provided lipid identifier. Therefore, enrichment based on double bond positions (location and conformation), bond type (ester and ether), and modifications (oxidized and cyclized) can be calculated. Human adipose tissue lipidomes (DEMO 1 and DEMO 2) demonstrate how LORA mines the information from widely used LC–MS lipidomic pipelines.<sup>13,14</sup> The human lung lipidome data set in DEMO 3 comes from Lipid Mini-On test files. The technical DEMO 4 data set contains a collection of (modified) lipids defined up to "Complete structure" to illustrate the potential use.

**Shared Structural Characteristics of Over-Represented Lipids.** The idea behind ORA is that we can infer a smaller set of structural characteristics for a set of significantly altered lipids, which reduces the dimensionality and defines the essential features of the query data. However, we can also take the sets of structural characteristics, explore their intersections, and find the lipid molecule(s) that best represent the set of significantly altered lipids. The lipids present in the most over-represented terms (lipids with the highest number of term intersections) are defined as Very Important Lipids (VILs).

Table 1 and Figure 2 show the analysis of DEMO 4. Table 1 defines the seven over-represented terms, and the UpSet plot in Figure 2 highlights the common patterns. The green cluster defines 13 lipids that share structural features in four over-represented terms (Figure 2 B and C). For example, one output is "The query lipids are statistically significantly enriched in glycerophospholipids containing 20:4 acyl at the *sn*-2 position".

**UpSet plot Visualization of Term Intersections.** The major challenge in evaluating the over-represented structural characteristics is the enormous number of set intersections if the number of sets exceeds a reasonable threshold. In the case of one to three sets, the Venn or Euler diagrams, which visualize all possible logical relations between the sets, can be used. In the case of two to 20 sets, the UpSet plot provides a comprehensive visualization.<sup>20</sup> The UpSet plot helps identify the main structural features of enriched lipids and highlights the VILs in a graphical representation (Figure 2A). Connected black dots represent the intersection of the terms labeled on top. The term intersection size (cardinality bar plot) represents the number of lipids that have this specific set of structural features in common. The *p*-values belong to the particular lipids within individual terms (*n* lipids  $\times$  *m* terms). The bar graph at the bottom shows how many lipids belong to the



term. The plot is interactive, and a table showing all lipids within the specific intersection (Figure 2C) is generated upon clicking on the cardinality bar.

**Hierarchical Tree Visualization of the Lipidome.** Goslin provides information on the hierarchical level of the lipids in the data set. We can visualize the information as a circular tree map (dendrogram) showing part-to-whole relationships and the level of structural details provided by the analytical method. The graph itself is interactive, and each lipid level can be explored via a tooltip (Figure S1A). The lipidome enrichment down to the LIPID MAPS SUBCLASS level can be visualized as an interactive network in Cytoscape, which allows for the mapping of statistical data onto nodes and edges (Figure S1B).

**Report.** LORA provides output in a zip archive containing a PDF report, parameter settings, over-representation analysis results, VIL table, UpSet plot, intersection tables, circular tree map, and the lipidome network in SVG, JPG, phyloXML, interactive HTML, XLSX, and Cytoscape formats, respectively.

**Limitations.** Lipid modifications, including oxidation, nitration, or halogenation, represent a new level of nomenclature complexity that extremely expands the search space.<sup>7</sup> Testing all structural characteristics up to “Complete structure level” would require computational resources that would not balance information gain. Therefore, the application does not implement “rare” features like the position of lipid modification or enantiomers. Lipid names in conflict with the hierarchical structure of shorthand notation (e.g., PC 16:1(7)\_16:1(9)) will be converted to the closest valid level (e.g., PC 16:1\_16:1). Goslin grammars (version 2) are currently limited to the most common set of lipid classes and do not consider multiple class categorization of lipids in LIPID MAPS (e.g., [FA01] Fatty Acid Conjugates and [FA02] Octadecanoids).<sup>11</sup>

## DISCUSSION AND CONCLUSION

Lipidomic analyses are usually performed to gain insight into system lipid metabolism, and ORA helps reduce the observation's dimensionality. In contrast to transcriptomics, where the ORA results can be directly used in pathways analysis,<sup>19</sup> functional pathway schemes for lipids are largely unavailable. The major problems are (1) the inability to assign a lipid database identifier to experimentally generated information describing the lipid molecule at the level of enzyme–substrate specificity,<sup>6</sup> (2) the continuous remodeling of the head groups and acyl chains by many enzymes simultaneously, and (3) the lack of curated lipid pathways at various nomenclature levels (with a few exceptions<sup>26,27</sup>). Generalized databases such as KEGG do not consider the structural diversity of lipids and often blur together multiple biological processes, thus compromising the biological interpretation. To overcome the problems, LORA builds on the Goslin standardization approach and known lipid structural characteristics provided by novel analytical techniques. Instead of relying on predefined schemes, LORA creates a set of over-represented terms based solely on provided structural information and statistical tests. The user can either directly interpret this set or reshape it by the UpSet plot to highlight the most common structural features and their representatives (lipid species).

The most common set visualization approach—Venn diagrams—do not scale beyond three or four sets. The UpSet plot, in contrast, is well suited for the quantitative analysis of data with more than three sets. When more than

seven sets intersect, the advanced UpSet plots allowing aggregation and grouping should be used to reduce the dimensionality.<sup>20</sup> We optimized the UpSet plot implemented in LORA for common lipidomics data sets. We limited the visualization to at most 13 terms because the calculation costs of all possible term intersections grow exponentially. Of note, the choice of lipid structural features, similar to the pathway database, used in ORA can have a much stronger effect on the enrichment results than the statistical corrections used in these analyses.<sup>19</sup>

LORA is a tool focused on the expanding technologies in (epi)lipidomics that allow for more precise identification of lipid structures. Routine use of supercritical fluid chromatography, ion mobility spectrometry, ion–molecule reactions, or derivatization techniques to specifically target double bond positions will provide further levels of detail to lead toward the full structural characterization of lipids. LORA mines this information-rich data set and helps interpret lipid structural features and over-represented terms using visualization tools. It is the next step toward understanding lipidomic data sets at the molecular level.

## ASSOCIATED CONTENT


### Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.analchem.3c02039>.

Supplementary figures and texts (PDF)


## AUTHOR INFORMATION


### Corresponding Author

Ondrej Kuda – *Institute of Physiology, Czech Academy of Sciences, 14220 Prague, Czechia*;  [orcid.org/0000-0001-7034-4536](https://orcid.org/0000-0001-7034-4536); Email: [ondrej.kuda@fgu.cas.cz](mailto:ondrej.kuda@fgu.cas.cz)

### Authors

Michaela Vondrackova – *Institute of Physiology, Czech Academy of Sciences, 14220 Prague, Czechia*

Dominik Kopczynski – *Institute of Analytical Chemistry, University of Vienna, 1090 Vienna, Austria*;  [orcid.org/0000-0001-5885-4568](https://orcid.org/0000-0001-5885-4568)

Nils Hoffmann – *Forschungszentrum Jülich, Institute of Bio- and Geosciences (IBG-5), 52428 Jülich, Germany*;  [orcid.org/0000-0002-6540-6875](https://orcid.org/0000-0002-6540-6875)

Complete contact information is available at:

<https://pubs.acs.org/doi/10.1021/acs.analchem.3c02039>

### Author Contributions

Conceptualization: O.K. Data curation: M.V. and O.K. Formal analysis: M.V., D.K., and N.H. Funding acquisition: O.K. Investigation: M.V. and O.K. Methodology: M.V. and O.K. Project administration: O.K. Resources: D.K., N.H., and O.K. Software: M.V., D.K., N.H., and O.K. Supervision: O.K. Validation: D.K. and N.H. Visualization: M.V. and O.K. Writing—original draft: M.V. and O.K. Writing—review and editing: M.V., N.H., and O.K.

### Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENTS

This work is supported by the project National Institute for Research of Metabolic and Cardiovascular Diseases (Programme EXCELES, ID Project No. LX22NPO5104) –

Funded by the European Union – Next Generation EU, Ministry of Health [NV19-02-00118], the Czech Academy of Sciences [Lumina Quaeruntur LQ200111901], and COST Action CA19105 - Pan-European Network in Lipidomics and EpiLipidomics (EpiLipidNET), supported by COST (European Cooperation in Science and Technology).

## REFERENCES

- (1) Šála, M.; Lisa, M.; Campbell, J. L.; Holčapek, M. *Rapid Commun. Mass Spectrom.* **2016**, *30*, 256–264.
- (2) Menzel, J. P.; Young, R. S. E.; Benfield, A. H.; Scott, J. S.; Wongsomboon, P.; Cudman, L.; Cvacka, J.; Butler, L. M.; Henriques, S. T.; Poad, B. L. J.; Blanksby, S. J. *Nat. Commun.* **2023**, *14*, 3940.
- (3) Ren, H.; Triebl, A.; Muralidharan, S.; Wenk, M. R.; Xia, Y.; Torta, F. *Analyst* **2021**, *146*, 3899–3907.
- (4) Brezinova, M.; Kuda, O.; Hansikova, J.; Rombaldova, M.; Balas, L.; Bardova, K.; Durand, T.; Rossmeisl, M.; Cerna, M.; Stranak, Z.; Kopecky, J. *BBA MCB* **2018**, *1863*, 126–131.
- (5) Hancock, S. E.; Poad, B. L.; Batarseh, A.; Abbott, S. K.; Mitchell, T. W. *Anal. Biochem.* **2017**, *524*, 45–55.
- (6) Liebisch, G.; Fahy, E.; Aoki, J.; Dennis, E. A.; Durand, T.; Ejsing, C. S.; Fedorova, M.; Feussner, I.; Griffiths, W. J.; Kofeler, H.; Merrill, A. H., Jr; Murphy, R. C.; O'Donnell, V. B.; Oskolkova, O.; Subramaniam, S.; Wakelam, M. J. O.; Spener, F. *J. Lipid Res.* **2020**, *61*, 1539–1555.
- (7) Damiani, T.; Bonciarelli, S.; Thallinger, G. G.; Koehler, N.; Krettler, C. A.; Salihoglu, A. K.; Korf, A.; Pauling, J. K.; Pluskal, T.; Ni, Z.; Goracci, L. *Anal. Chem.* **2023**, *95*, 287–303.
- (8) Molenaar, M. R.; Jeucken, A.; Wassenaar, T. A.; van de Lest, C. H. A.; Brouwers, J. F.; Helms, J. B. *Gigascience* **2019**, *8*, giz061.
- (9) Clair, G.; Reehl, S.; Stratton, K. G.; Monroe, M. E.; Tfaily, M. M.; Ansong, C.; Kyle, J. E. *Bioinformatics* **2019**, *35*, 4507–4508.
- (10) Kopczynski, D.; Hoffmann, N.; Peng, B.; Ahrends, R. *Anal. Chem.* **2020**, *92*, 10957–10960.
- (11) Kopczynski, D.; Hoffmann, N.; Peng, B.; Liebisch, G.; Spener, F.; Ahrends, R. *Anal. Chem.* **2022**, *94*, 6097–6101.
- (12) Ni, Z.; Fedorova, M. LipidLynxX: a data transfer hub to support integration of large scale lipidomics datasets. *bioRxiv Preprint*, 2020. DOI: 10.1101/2020.04.09.033894.
- (13) Janovska, P.; Melenovsky, V.; Svobodova, M.; Havlenova, T.; Kratochvilova, H.; Haluzik, M.; Hoskova, E.; Pelikanova, T.; Kautzner, J.; Monzo, L.; Jurcova, I.; Adamcova, K.; Lenkova, L.; Buresova, J.; Rossmeisl, M.; Kuda, O.; Cajka, T.; Kopecky, J. *J. Cachexia Sarcopenia Muscle* **2020**, *11*, 1614–1627.
- (14) Lange, M.; Angelidou, G.; Ni, Z.; Criscuolo, A.; Schiller, J.; Blüher, M.; Fedorova, M. *Cell Rep. Med.* **2021**, *2*, 100407.
- (15) Lauder, S. N.; Allen-Redpath, K.; Slatter, D. A.; Aldrovandi, M.; O'Connor, A.; Farewell, D.; Percy, C. L.; Molhoek, J. E.; Rannikko, S.; Tyrrell, V. J.; Ferla, S.; Milne, G. L.; Poole, A. W.; Thomas, C. P.; Obaji, S.; Taylor, P. R.; Jones, S. A.; de Groot, P. G.; Urbanus, R. T.; Horkko, S.; Uderhardt, S.; Ackermann, J.; Vince Jenkins, P.; Branciale, A.; Kronke, G.; Collins, P. W.; O'Donnell, V. B. *Sci. Signal* **2017**, *10*, ean2787.
- (16) Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S. J.; Brett, M.; Wilson, J.; Millman, K. J.; Mayorov, N.; Nelson, A. R. J.; Jones, E.; Kern, R.; Larson, E.; Carey, C. J.; et al. *Nat. Methods* **2020**, *17*, 261–272.
- (17) Chen, S. Y.; Feng, Z.; Yi, X. *J. Thorac. Dis.* **2017**, *9*, 1725–1729.
- (18) Huang, D. W.; Sherman, B. T.; Lempicki, R. A. *Nucleic Acids Res.* **2009**, *37*, 1–13.
- (19) Karp, P. D.; Midford, P. E.; Caspi, R.; Khodursky, A. *BMC Genomics* **2021**, *22*, 191.
- (20) Lex, A.; Gehlenborg, N.; Strobel, H.; Vuilleumot, R.; Pfister, H. *IEEE Trans Vis Comput. Graph* **2014**, *20*, 1983–1992.
- (21) Cock, P. J.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; de Hoon, M. J. *Bioinformatics* **2009**, *25*, 1422–1423.
- (22) Han, M. V.; Zmasek, C. M. *BMC Bioinformatics* **2009**, *10*, 356.
- (23) Lopes, M.; Brejchova, K.; Riecan, M.; Novakova, M.; Rossmeisl, M.; Cajka, T.; Kuda, O. *Cell Rep.* **2021**, *37*, 109833.
- (24) Van Rossum, G.; Drake, F. L. *Python 3 Reference Manual*; CreateSpace: Scotts Valley, CA, 2009.
- (25) Sievert, C. *Interactive Web-Based Data Visualization with R, plotly, and shiny*; Chapman and Hall/CRC: FL, 2020.
- (26) Martens, M.; Ammar, A.; Riutta, A.; Waagmeester, A.; Slenter, D. N.; Hanspers, K.; Miller, R. A.; Digles, D.; Lopes, E. N.; Ehrhart, F.; Dupuis, L. J.; Winckers, L. A.; Coort, S. L.; Willighagen, E. L.; Evelo, C. T.; Pico, A. R.; Kutmon, M. *Nucleic Acids Res.* **2021**, *49*, D613–D621.
- (27) Gaud, C.; Sousa, B. C.; Nguyen, A.; Fedorova, M.; Ni, Z.; O'Donnell, V. B.; Wakelam, M. J. O.; Andrews, S.; Lopez-Clavijo, A. F. *FI000Res.* **2021**, *10*, 4.