Data in Brief 35 (2021) 106822

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

A comprehensive dataset of the extra virgin olive oil (EVOO) proteome

Antonio Jesús Castro, Elena Lima-Cabello, Juan de Dios Alché*

Plant Reproductive Biology and Advanced Imaging Laboratory, Department of Biochemistry, Cell and Molecular Biology of Plants, Estación Experimental del Zaidín (CSIC), 18008 Granada, Spain

ARTICLE INFO

Article history:
Received 25 January 2021
Accepted 28 January 2021
Available online 30 January 2021

Keywords:
Extra virgin olive oil (EVOO)
Lipoxygenase
Olea europaea
Proteomics
seed storage proteins (SSP)

ABSTRACT

Proteins and peptides are minor components of vegetal oils. The presence of these compounds in virgin olive oil was first reported in 2001, but the nature of the olive oil proteome is still a puzzling question for food science researchers. In this paper, we have compiled for a first time a comprehensive proteomic dataset of olive fruit and fungal proteins that are present at low but measurable concentrations in a vegetable oil from a crop of great agronomical relevance as olive (*Olea europaea* L.). Accurate mass nLC-MS data were collected in high definition direct data analysis (HD-DDA) mode using the ion mobility separation step. Protein identification was performed using the Mascot Server v2.2.07 software (Matrix Science) against an ad hoc database made of olive protein entries. Starting from this proteomic record, the impact of these proteins on olive oil stability and quality could be tested. Moreover, the effect of olive oil proteins on human health and their potential use as functional food components could be also evaluated. In addition, this dataset provides a resource for use in further functional comparisons across other vegetable oils, and also expands the proteomic resources to non-model species, thus also allowing further comparative

* Corresponding author.
E-mail address: juandedios.alche@eez.csic.es (J.d.D. Alché).
Social media: (J.d.D. Alché)

https://doi.org/10.1016/j.dib.2021.106822

inter-species studies. The data presented here are related to the research article of Castro et al. [1].

Specifications Table

Subject	Agricultural and Biological Sciences
Specific subject area	Food Science, biochemical composition of vegetable oils
Type of data	Drop-drown tables, figures
How data were acquired	Data were acquired by nanoLC-MS using a nanoAcquity UPLC system (Waters Corp.) and a Synapt G2Si ESI Q-Mobility-TOF spectrometer (Waters Corp.) equipped with an ion mobility chamber (T-Wave-IMS)
Data format	Excel files with data analysis output, figures embedded in a single PDF file. Raw data are also available on a public data repository (for more information, please see the Data accessibility section below)
Parameters for data collection	Extra virgin olive (cv. Picual) oil (2018 harvest) of the Protected Designation of Origin (PDO) "Montes de Granada" was used as material for protein extraction
Description of data collection	Proteins were extracted from five liters of EVOO and subjected to SDS-PAGE separation prior to nLC-MS analysis. Accurate mass nLC-MS data were collected in high definition direct data analysis (HD-DDA) mode using the ion mobility separation step. Protein identification was performed using the Mascot Server v2.2.07 software (Matrix Science) against an in-house olive protein database
Data source location	Estación Experimental del Zaidín (CSIC), Granada, Spain
Data accessibility	The mass spectrometry raw data have been deposited to the ProteomeXchange Consortium via the PRIDE [2,3] partner repository with the dataset identifier PXD019894 (http://www.ebi.ac.uk/pride/archive/projects/PXD019894). Analyzed data are with this article.
Related research article	A.J. Castro, E. Lima-Cabello, J.D. Alché, Identification of seed storage proteins as the major constituents of the extra virgin olive oil proteome, Food Chemistry: X 7 (2020) 100,099 http://doi.org/10.1016/j.fochx.2020.100099

Value of the Data

- This proteomic dataset provides a comprehensive list of olive fruit and fungal proteins present in the extra virgin olive oil (EVOO).
- Starting from these proteomic data, the impact of different protein components present in EVOO on its stability and quality could be further studied.
- The effect of the olive oil proteins on human health and their potential use as functional food components could be also evaluated.
- This dataset provides a resource for use in further functional comparisons across other vegetable oils.
- Expanding the proteomic resources to non-model but agronomically relevant species will allow further comparative inter-species studies.

1. Data Description

A comprehensive dataset of olive fruit (i.e. pulp and seed) proteins identified in the extra virgin olive (cv. Picual) oil (EVOO) is provided in Table S1. In addition, Table S2 provides a list

of fungal proteins present in EVOO, derived from the yeast-like fungus *Aureobasidium pullulans*, which is ubiquitous in the phyllosphere and carposphere of the olive tree [4]. Tables S1 and S2 are available with this article as Excel (.xlsx) files and each list of identified proteins is displayed in drop-down table format. Proteins and organisms in which proteins were identified were annotated according to NCBI nr database [5]. Main tables also incorporate a number of parameters about each protein identified, including: a) the gel slice in which the protein was isolated and identified, b) the number of amino acid residues, c) the theoretical mass and isoelectric point, d) the total percent coverage (i.e. the number of AA in all found peptides divided by the total number of AA in the entire protein sequence), e) the total number of identified proteins in the protein group of a master protein, f) the total number of peptide sequences unique to a protein group, g) the total number of distinct peptide sequences in the protein group, and h) the total number of identified peptide sequences (PSM, peptide spectrum matches) for the protein. Mascot searches were carried out against olive genome (https://denovo.cnag.cat/olive; [6]) and transcriptome (http://reprolive.eez.csic.es/olivodb; [7]) records and the Uniprot database (https://www.uniprot.org/; [8]). For each search, the protein score (i.e. the sum of the score of the individual peptides), the protein coverage, the number of distinct peptide sequences in the protein group and the number of identified peptide sequences are also provided.

For each protein identified, the sheet can be expanded by clicking on the [+] key located on the left margin of the Excel file, which opens the row parameters for the associated peptides, including: a) the AA sequences of identified peptides, b) the number of PSMs for the protein, c) the number of proteins in which this peptide is found, d) the number of protein groups in which this peptide is found, e) fixed and variable modifications of peptides, f) the number of missed cleavage positions, g) the monoisotopic mass of the peptide, h) the accession number of the protein in the corresponding database, i) the top level confidence (only the high-confidence data were considered) achieved with the peptide sequence, j) the score for the peptide after MASCOT search in the corresponding database, and k) the experimental *m/z* value for the peptide after MASCOT search in the corresponding database. MS/MS spectra corresponding to hand-validated peptides are also included as Figs. S1–S23 embedded in a single PDF file. The mass spectrometry raw data were deposited to the ProteomeXchange Consortium via the PRIDE [2,3] partner repository with the dataset identifier PXD019894 (www.ebi.ac.uk/pride/archive/projects/PXD019894). Alternatively, raw data files can be also downloaded from ftp://ftp.pride.ebi.ac.uk/pride/data/archive/2020/09/PXD019894.

2. Experimental Design, Materials and Methods

2.1. Materials

Freshly bottled extra virgin olive (cv. Picual) oil (2018 harvest) of the Protected Designation of Origin (PDO) “Montes de Granada” was purchased from a local market and stored in the dark at 15–18 °C until use. All chemicals used had purity greater than 99%.

2.2. In situ *digestion of extra virgin olive oil proteins*

Extra virgin olive oil proteins were extracted, electrophoresed on 1-D polyacrylamide gels and stained as described in [1]. The gel lane containing the EVOO proteins was systematically cut from the top (slice S1) to the bottom (slice S10) into slices of ~1 cm width each (see Fig. 1A in ref. [1]). *In situ* digestion of proteins was performed using the MassPREP Station (Micromass, Manchester, UK). Gel slices were washed three times in a mixture containing 25 mM NH_4HCO_3: acetonitrile (ACN) (1:1, v/v). The Cys-residues were reduced by 50 µL of 10 mM dithiothreitol (DTT) at 57 °C and alkylated by 50 µL of 55 mM iodoacetamide at room temperature. After gel dehydration with ACN, proteins were digested overnight at room temperature in 15 µL of a

solution containing 12.5 ng μL^{-1} of a modified porcine trypsin (Promega, Madison, WI, USA) prepared in 25 mM NH_4HCO_3. Finally, a double extraction was performed, first with 60% (v/v) ACN in 5% (v/v) formic acid, and subsequently with 100% (v/v) ACN.

2.3. LC-MS data acquisition and analysis

Nano-Liquid chromatography (nLC) of the resulting tryptic peptides was performed using a nanoACQUITY UPLC© system (Waters, Milford, MA, USA), equipped with a nanoACQUITY UPLC© Peptide BEH C_{18} (200 mm length × 75 µm ID, 1.7 µm particle size) nano-column (catalog no. 186,007,483, Waters), coupled with a nanoACQUITY UPLC C_{18} trap column (20 mm length × 180 µm ID, 5 µm particle size) (catalog no. 186,007,496, Waters). About 0.5 µg was loaded per run. The solvent system consisted of 0.1% (v/v) formic acid in water (mobile phase A) and 0.1% (v/v) formic acid in acetonitrile (ACN) (mobile phase B). Elution was performed at a flow rate of 300 nL min^{-1}, using a linear gradient (5 to 60%) of mobile phase B over a chromatographic ramp of 120 min. A lock mass compound [Glu1]-Fibrinopeptide B (100 fmol μl^{-1}) (catalog no. 196,007,091–2, Waters) was delivered by an auxiliary pump of the LC system at 500 nL min^{-1} to the reference sprayer of the NanoLockSpray Exact Mass Ionization source (Waters Corp.) of the mass spectrometer.

A Synapt G2Si ESI Q-Mobility-TOF spectrometer (Waters Corp.) equipped with an ion mobility chamber (T-Wave-IMS) was used for high definition data acquisition analysis. The mass spectrometer was operating in positive mode ESI with the following settings: source temperature was set to 120 °C, while dry gas flow was at 3.7 ml min^{-1}. The nano-electrospray voltage was optimized to 1.3 kV. Data were post-acquisition lock mass corrected using the double charged monoisotopic ion of [Glu1]-Fibrinopeptide B. Accurate LC-MS data were collected in High Definition Direct Data Analysis (HD-DDA) mode that enhances signal intensities using the ion mobility separation step [9].

2.4. Database search and protein identification

Protein identification was performed using the Mascot Server v2.2.07 software (Matrix Science, London, UK) against an ad hoc-generated database composed of protein entries retrieved from the olive genome [6] and transcriptome [7] records, as well as known contaminant proteins such as human keratins and trypsin, extracted from the NCBI nr protein database [5]. To be accepted for the identification, an error of less than 15 ppm of peptide mass tolerance and 0.2 Da of fragment mass tolerance were tolerated. Up to 3 missed cleavage points were allowed and some modifications were taken into account: carbamidomethylation of Cys-residues (+57 Da) as fixed modification, oxidation of Met (+16 Da) as variable modification, and peptide charges of +2 and +3. In addition, searches were performed without any molecular weight (Mr) or isoelectric point (pI) restrictions. To calculate the false discovery rate (FDR) [10], the search was performed using the “decoy” option in Mascot (Matrix Science). Peptide identifications extracted from Mascot result files were validated at a final peptide FDR of 1%. Peptide matches were also manually validated if their score was close to the Mascot homology threshold for a given Mascot p value.

CRediT Author Statement

Antonio Jesús Castro: Conceptualization, Investigation, Data curation, Visualization, Writing - Original draft preparation, Writing - Reviewing and Editing; **Elena Cabello-Lima:** Investigation, Validation, Writing - Reviewing and Editing; **Juan de Dios Alché:** Conceptualization, Funding Acquisition, Supervision, Writing - Reviewing and Editing.

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 has been supported by FEDER-cofinanced grants RTC-2017-6654-2 (MINECO), AT17_5247 and P18-RT-1577 (Junta de Andalucía), and the Technological Agreement no. 09021200008. The authors thank José Berral and Dr. Adoración Zafra for their technical assistance. Mass spectrometry analysis was carried out at the CIC-BioGUNE's Proteomic Platform in Derio, Bizkaia, Spain. CIC-bioGUNE is part of BRTA (Basque Research and Technology Alliance) and is supported by ProteoRed-ISCIII (grant PRB3 IPT17/0019), CIBERehd Network and Severo Ochoa grant (SEV-2016-0644).

Supplementary Materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.dib.2021.106822.

References

[1] A.J. Castro, E. Lima-Cabello, J.D. Alché, Identification of seed storage proteins as the major constituents of the extra virgin olive oil proteome, Food Chem. X 7 (2020), doi:10.1016/j.fochx.2020.100099.
[2] Y. Perez-Riverol, A. Csordas, J. Bai, M. Bernal-Llinares, S. Hewapathirana, D.J. Kundu, A. Inuganti, J. Griss, G. Mayer, M. Eisenacher, E. Pérez, J. Uszkoreit, J. Pfeuffer, T. Sachsenberg, S. Yilmaz, S. Tiwary, J. Cox, E. Audain, M. Walzer, A.F. Jarnuczak, T. Ternent, A. Brazma, J.A. Vizcaíno, The PRIDE database and related tools and resources in 2019: improving support for quantification data, Nucl. Acids Res. 47 (2019) D442–D450, doi:10.1093/nar/gky1106.
[3] E.W. Deutsch, N. Bandeira, V. Sharma, Y. Perez-Riverol, J.J. Carver, D.J. Kundu, D. García-Seisdedos, A.F. Jarnuczak, S. Hewapathirana, B.S. Pullman, J. Wertz, Z. Sun, S. Kawano, S. Okuda, Y. Watanabe, H. Hermjakob, B. MacLean, M.J. MacCoss, Y. Zhu, Y. Ishihama, J.A. Vizcaíno, The ProteomeXchange consortium in 2020: enabling 'big data' approaches in proteomics, Nucl. Acids Res. 48 (2020) D1145–D1152, doi:10.1093/nar/gkz984.
[4] A. Abdelfattah, M.G.L.D. Nicosia, S.O. Cacciola, S. Droby, L Schena, Metabarcoding analysis of fungal diversity in the phyllosphere and carposphere of olive (Olea europaea), PLoS ONE 10 (2015) e0131069, doi:10.1371/journal.pone.0131069.
[5] NCBI nr database. https://www.ncbi.nlm.nih.gov/refseq/about/nonredundantproteins/, (Accessed 17 May 2020).
[6] R. Carmona, A. Zafra, P. Seoane, A.J. Castro, D. Guerrero-Fernández, T. Castillo-Castillo, A. Medina-García, F.M. Cánovas, J.F. Aldana-Montes, I. Navas-Delgado, J.D. Alché, M.G. Claros, ReprOlive: a database with linked data for the olive tree (Olea europaea L.) reproductive transcriptome. Front, Plant Sci. 6 (2015) 625, doi:10.3389/fpls.2015.00625.
[7] F. Cruz, I. Julca, J. Gómez-Garrido, D. Loska, M. Marcet-Houben, E. Cano, B. Galán, L. Frias, P. Ribeca, S. Derdak, M. Gut, M. Sánchez-Fernández, J.L. García, I.G. Gut, P. Vargas, T.S. Alioto, T. Gabaldón, Genome sequence of the olive tree, Olea europaea, GigaSci 5 (2016) 29, doi:10.1186/s13742-016-0134-5.
[8] The UniProt Consortium, UniProt: a worldwide hub of protein knowledge, Nucl. Acids Res. 47 (2019) D506–D515, doi:10.1093/nar/gky1049.
[9] D. Helm, J.P.C. Vissers, C.J. Hughes, H. Hahne, B. Ruprecht, F. Pachl, A. Grzyb, K. Richardson, J. Wildgoose, S.K. Maier, H. Marx, M. Wilhelm, I. Becher, S. Lemeer, M. Bantscheff, J.I. Langridge, B. Kuster, Ion mobility tandem mass spectrometry enhances performance of bottom-up proteomics, Mol. Cell. Proteom. 13 (2014) 3709–3715, doi:10.1074/mcp.M114.041038.
[10] J.E. Elias, S.P. Gygi, Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry, Nat. Methods 4 (2007) 207–214, doi:10.1038/nmeth1019.