Data in Brief 13 (2017) 124–128

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

Data from proteome analysis of *Lasiodiplodia theobromae* (Botryosphaeriaceae)

Carla C. Uranga [a], Majid Ghassemian [b], Rufina Hernández-Martínez [a]

[a] *Centro de Investigación Científica y de Educación Superior de Ensenada (CICESE), Carretera Ensenada-Tijuana 3918, Zona Playitas, 22860 Ensenada, B.C., Mexico*
[b] *University of California, San Diego, Department of Chemistry and Biochemistry, 9500 Gilman Drive, La Jolla, CA 92093-0378, United States*

ARTICLE INFO

Article history:
Received 10 March 2017
Accepted 27 April 2017
Available online 17 May 2017

Keywords:
Trunk-disease fungi
LC-nanoESI-MS
Proteomics
de novo peptide sequencing
Gene ontology

ABSTRACT

Trunk disease fungi are a global problem affecting many economically important fruiting trees. The Botryosphaeriaceae are a family of trunk disease fungi that require detailed biochemical characterization in order to gain insight into their pathogenicity. The application of a modified Folch extraction to protein extraction from the Botryosphaeriaceae *Lasiodiplodia theobromae* generated an unprecedented data set of protein identifications from fragmentation analysis and *de novo* peptide sequencing of its proteome. This article contains data from protein identifications obtained from a database-dependent fragmentation analysis using three different proteomics algorithms (MSGF, Comet and X! Tandem *via* the SearchGUI proteomics pipeline program) and *de novo* peptide sequencing. Included are data sets of gene ontology annotations using an all-Uniprot ontology database, as well as a *Saccharomyces cerevisiae*-only and a *Candida albicans*-only ontology database, in order to discern between those proteins involved in common functions with *S. cerevisiae* and those in common with the pathogenic yeast *C. albicans*. Our results reveal the proteome of *L. theobromae* contains more ontological categories in common to *C. albicans*, yet possesses a much wider metabolic repertoire than any of the yeasts studied in this work. Many novel proteins of interest were identified for further biochemical characterization and annotation efforts, as further discussed in the article referencing this article (1). Interactive Cytoscape networks of molecular

DOI of original article: http://dx.doi.org/10.1016/j.biopen.2017.03.001

E-mail addresses: urangacarla@gmail.com (C.C. Uranga), mghassemian@ucsd.edu (M. Ghassemian), ruhernan@cicese.mx (R. Hernández-Martínez).

http://dx.doi.org/10.1016/j.dib.2017.04.058

functions of identified peptides using an all-Uniprot ontological database are included. Data, including raw data, are available *via* ProteomeXchange with identifier PXD005283.

Specifications Table

Subject area	*Database-dependent peptide fragmentation protein identification and de novo sequencing of peptides from filamentous fungi*
More specific subject area	*Database-dependent peptide fragmentation-based protein identifications and de novo peptide sequencing of peptides from Lasiodiplodia theobromae*
Type of data	*SearchGUI protein identification data and DeNovoGUI peptide sequencing data*
How data was acquired	*Mass spectrometry with an AB SCIEX Triple TOF mass spectrometer*
Data format	*Protein identifications were analyzed and filtered with a $<1\%$ False Discovery Rate (FDR). De novo sequencing data reports the most homologous peptides in the Uniprot database to those from L. theobromae.*
Experimental factors	*Folch Extraction of triplicate incubations of L. theobromae in Vogel's salts supplemented with both 5% glucose and 5% grapeseed oil.*
Experimental features	*From previous work, it was shown that under these conditions, fatty acid esters with physiological activity in plants were detected and produced abundantly [2].*
Data source location	*Ensenada, Baja California, Mexico and San Diego, California, USA.*
Data accessibility	*The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE [3] partner repository with the dataset identifier* PXD005283.

Value of the data

- This is the first report of peptide fragmentation and *de novo* sequencing analysis of *Lasiodiplodia theobromae*, a pathogen of primarily economically important fruiting trees.
- Peptide fragmentation and *de novo* peptide sequencing analysis yielded many novel protein identifications that may aid in understanding fungal metabolism and further advance biochemical annotation efforts.
- The application of a Folch extraction is novel for filamentous fungi, and resulted in an extensive data set that may makes future proteomics experiments more efficient.
- Bioinformatics of filamentous fungi requires development, and this data is a step in that direction. Cytoscape networks are included for molecular function annotations for database-dependent protein identifications *via* fragmentation analysis and *de novo* peptide sequencing.

1. Data

The data consists in database-dependent peptide fragmentation analysis using the SearchGUI program [2] with the MSGF, Comet, and X!Tandem sequencing algorithms, identified with a 1% FDR. The entire list of protein identifications is available in this article as Supplementary data. Interactive Cytoscape networks are included, as well as full ontology reports using an all-Uniprot annotation database, as well as *Saccharomyces cerevisiae*-only and *Candida albicans*-only annotation databases. *De* *novo* peptide sequencing results are included in Supplementary data, as well as complete BLASTp protein identification results.

2. Experimental design, materials and methods

L. theobromae UCD256Ma (isolated in Madera County, California, USA) was provided by Dr. Douglas Gubler from the University of California at Davis [3], and incubated and extracted as described in the accompanying article [1]. Briefly, 0.5 g of the solids from the 50 mL fungal incubations (*L. theobromae* incubated in 5% glucose and 5% grapeseed oil and Vogel's salts for 20 days and lyophilized) remaining from the Folch extraction were dried under a stream of nitrogen and re-suspended in 50 mM Tris buffer, pH 8.00. Acetonitrile was added to the sample to a final concentration of 10%. The samples were then boiled for 5 min and cooled to room temperature. TCEP (Tris (2-carboxyethyl) phosphine) was added to 1 mM (final concentration) and the samples were incubated at 37 °C for 30 min. Subsequently, the samples were carboxymethylated with 0.5 mg/ml of iodoacetamide for 30 min at 37 °C in dark followed by neutralization with 2 mM TCEP (final concentration). Samples were boiled for 10 min followed by protease digestion with a 1:100 ratio of trypsin: protein (Pierce™ Trypsin Protease, MS Grade Catalog number: 90057 with K, R specificity). After an overnight digestion, samples were centrifuged on a desktop microfuge at max speed (15,000 rpm) for 10 min to remove the insoluble fraction. The soluble fraction was adjusted to 0.2% formic acid and 5% acetonitrile and its peptide content isolated using C-18 solid phase extraction (Thermo Scientific, PI-87782) as described by the manufacturer.

Proteomics mass spectrometry was done at the University of California, San Diego as described in the accompanying article [1]. The nano-spray ionization experiments were performed using a TripleTOF 5600 hybrid mass spectrometer (ABSCIEX) interfaced with a nano-scale reversed-phase UPLC (Waters nano ACQUITY) using a 20 cm to 75 µM ID glass capillary packed with 2.5-µM C18 (130) CSH™ beads (Waters). Peptides were eluted from the C18 column into the mass spectrometer with a linear gradient (5–80%) of acetonitrile (ACN) at a flow rate of 250 µL/min for 90 min. The buffers used to create the ACN gradient were Buffer A (98% H_2O, 2% ACN, 0.1% formic acid and 0.005% TFA) and Buffer B (100% ACN, 0.1% formic acid, and 0.005% TFA). MS/MS data were obtained in a data-dependent manner in which the MS1 data was acquired for 250 ms at m/z of 400–1250 Da and the MS/MS data was acquired from m/z of 50 to 2000 Da. An MS1-TOF acquisition time of 250 ms was set, followed by 50 MS2 events of 48 ms acquisition time for each event. The threshold to trigger the MS2 event was set to 150 counts, when the ion had the charge state +2, +3 and +4. The ion exclusion time was set to 4 s.

3. Protein identification

This information appears as in the accompanying article [1]. This methodology is replicated in this article for the reader's convenience. Peak lists obtained from MS/MS spectra were identified using X! Tandem Vengeance (2015.12.15.2) [4], MS-GF+ version Beta (v10282) [5] and either OMSSA version 2.1.9 [6] or Comet version 2016.01 rev. 2 [7]. The search was conducted using SearchGUI version 3.1.2 [2]. The data was searched against a whole Uniprot/Swissprot database search (manually annotated and reviewed), [8] as well as a non-redundant Botryosphaeriaceae-only database downloaded from NCBI [9]. An all-human database from Uniprot was also used for further assessing protein identifications. All identification data from each database may be found as Supplementary Data S2, S3 and S5.

The identification settings were as follows: Trypsin with a maximum of 2 missed cleavages; 60.0 ppm as MS1 and 0.8 Da as MS2 tolerances; fixed modifications: Carbamidomethylation of C (+57.021464 Da) and Oxidation of M (+15.994915 Da), variable modifications: Acetylation of protein N-term (+42.010565 Da), Pyrolidone from E (+18.010565 Da), Pyrolidone from Q (+17.026549 Da) and Pyrolidone from carbamidomethylated C (+17.026549 Da). All algorithm- specific settings are listed in the Certificate of Analysis available in Supplementary Data S1.

Peptides and proteins were inferred from the spectrum identification results using PeptideShaker version 1.13.6 [10]. Peptide Spectrum Matches (PSMs), peptides and proteins were validated at a 1.0%

False Discovery Rate (FDR) estimated using a decoy hit distribution. All validation thresholds are listed in the Certificate of Analysis and are available in Supplementary Data S1A, S1B, and S1C and for all databases searched. Post-translational modification localizations were scored using the D-score [11] and the A-score [12] with a threshold of 95.0 as implemented in the compomics-utilities package [13]. An A-score above 95.0 was considered a confident localization. The mass spectrometry data along with the identification results have been deposited to the ProteomeXchange Consortium [14] *via* the PRIDE partner repository [15] with the dataset identifier PXD005283.

Gene ontology (GO) analysis of enriched proteins was done on all those hits obtained from the Uniprot database [8]. The software Cytoscape [16] with the BiNGO plugin [17] was used for GO and enrichment analysis using up-to-date databases, applying a hypergeometric test with a significance level (p-value) < 0.05, as well as a Benjamini and Hochberg false discovery rate (FDR) correction. Interactive Cytoscape BiNGO networks were created with data from all algorithms, and annotated with an all-Uniprot ontology database, available in Fig. 1 in this article. All gene ontology annotations may be found in Supplementary Data S5 and S6 in this article.

De novo peptide sequencing was performed in order to compare results and explore peptides *via* sequence homology with sequenced proteins found in the entire Uniprot database using BLAST. The program DeNovoGUI version 1.14.5 was used for this purpose [18], and both Novor [19] and PepNovo [20] were used for peptide sequencing. The mass allowance parameters were, for precursor mass tolerance: 10 ppm, and a fragment mass tolerance of 0.5 Da. Post-translational modification settings consisted in carbamidomethylation of cysteine (fixed) and oxidation of methionine (variable). All peptides were searched against the entire Uniprot database using a standalone version of NCBI-BLAST [21], with one peptide match per spectrum (most significant) and one BLAST match per peptide (most significant, lowest E-value). The BLAST match data was also analyzed for gene ontology (molecular functions) as described.

Acknowledgments

This work was made possible by a doctoral stipend from CONACyT for Carla Uranga, as well as support from the UCMEXUS Program, and consists in part of the doctoral thesis by Carla Uranga. Thanks to Dr. Elizabeth Komives from the UCSD proteomics mass spectrometry department for UCMEXUS support and sponsorship. Special thanks to Dr. Kristina Herbert and Dr. Genaro Pimienta for "omica" server establishment and access at CICESE. We are grateful to Dr. Marc Vaudel and Dr. Harald Barsnes for help with the SearchGUI program. Thanks to Claudio Espinosa de los Monteros for help with Adobe Photoshop.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.dib.2017.04.058.

References

[1] C.C. Uranga, M. Ghassemian, R. Hernandez-Martinez, Novel proteins from proteomic analysis of the trunk disease fungus Lasiodiplodia theobromae (Botryosphaeriaceae), Biochim. Open (2017).

[2] M. Vaudel, H. Barsnes, F.S. Berven, A. Sickmann, L. Martens, SearchGUI: an open-source graphical user interface for simultaneous OMSSA and X!Tandem searches, Proteomics 11 (5) (2011) 996–999.

[3] J.R. Úrbez-Torres, G.M. Leavitt, T.M. Voegel, W.D. Gubler, Identification and distribution of Botryosphaeria spp. Associated with grapevine cankers in California, Plant Dis. 90 (12) (2006) 1490–1503.

[4] R. Craig, R.C. Beavis, TANDEM: matching proteins with tandem mass spectra, Bioinformatics 20 (9) (2004) 1466–1467.

[5] S. Kim, P.A. Pevzner, MS-GF+ makes progress towards a universal database search tool for proteomics, Nat. Commun. [Internet] Nat. Publ. Group 5 (2014) 5277 (Available from) http://www.nature.com/doifinder/10.1038/ncomms6277.

[6] L.Y. Geer, S.P. Markey, J.A. Kowalak, L. Wagner, M. Xu, D.M. Maynard, et al., Open mass spectrometry search algorithm research articles, J. Proteome Res. (2004) 1–25.

[7] J.K. Eng, T.A. Jahan, M.R. Hoopmann, Comet: an open-source MS/MS sequence database search tool, Proteomics 13 (1) (2013) 22–24.

[8] A. Bateman, M.J. Martin, C. O'Donovan, M. Magrane, R. Apweiler, E. Alpi, et al., UniProt: a hub for protein information, Nucleic Acids Res. 43 (D1) (2015) D204–D212.

[9] L.Y. Geer, A. Marchler-Bauer, R.C. Geer, L. Han, J. He, S. He, et al., The NCBI BioSystems database, Nucleic Acids Res 38 (Suppl. 1) (2009) S492–S496.

[10] M. Vaudel, J.M. Burkhart, R.P. Zahedi, E. Oveland, F.S. Berven, A. Sickmann, et al., PeptideShaker enables reanalysis of MS-derived proteomics data sets, Nat. Biotechnol. 33 (1) (2015) 22–24.

[11] M. Vaudel, D. Breiter, F. Beck, J. Rahnenführer, L. Martens, R.P. Zahedi, D-score: a search engine independent MD-score, Proteomics 13 (6) (2013) 1036–1041.

[12] S.A. Beausoleil, J. Villén, S.A. Gerber, J. Rush, S.P. Gygi, A probability-based approach for high-throughput protein phosphorylation analysis and site localization, Nat. Biotechnol. 24 (10) (2006) 1285–1292.

[13] H. Barsnes, M. Vaudel, N. Colaert, K. Helsens, A. Sickmann, F.S. Berven, et al., compomics-utilities: an open-source Java library for computational proteomics, BMC Bioinform. [Internet] Biomed. Cent. Ltd. 12 (March) (2011) 70 (Available from) http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3060842&tool=pmcentrez&rendertype=abstract.

[14] J. Vizcaíno, E.E.W. Deutsch, R. Wang, J.A. Vizcaino, E.E.W. Deutsch, R. Wang, et al., ProteomeXchange provides globally coordinated proteomics data submission and dissemination, Nat. Biotechnol. 32 (3) (2014) 223–226. http://dx.doi.org/10.1038/nbt.2839 ⟨http://www.nature.com/nbt/journal/v32/n3/abs/nbt.2839.html#supplementary-information⟩, ⟨http://www.nature.com/doifinder/10.1038/nbt.2839⟩, ⟨http://www.nature.com/nbt/journal/v32/n3/full/nbt.2839.html⟩.

[15] L. Martens, H. Hermjakob, P. Jones, M. Adamsk, C. Taylor, D. States, et al., PRIDE: the proteomics identifications database, Proteomics 5 (13) (2005) 3537–3545.

[16] R. Christmas, I. Avila-Campillo, H. Bolouri, B. Schwikowski, M. Anderson, R. Kelley, et al., Cytoscape: a software environment for integrated models of biomolecular interaction networks, Am. Assoc. Cancer Res Educ. B (Karp 2001) (2005) 12–16.

[17] S. Maere, K. Heymans, M. Kuiper, BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks, Bioinformatics 21 (16) (2005) 3448–3449.

[18] T. Muth, L. Weilnböck, E. Rapp, C.G. Huber, L. Martens, M. Vaudel, et al., DeNovoGUI: an open source graphical user interface for de novo sequencing of tandem mass spectra, J. Proteome Res. 13 (2) (2014) 1143–1146.

[19] B. Ma, Novor: real-time peptide de Novo sequencing software, J Am. Soc. Mass Spectrom. 26 (11) (2015) 1885–1894.

[20] A. Frank, P. Pevzner, PepNovo: de novo peptide sequencing via probabilistic network modeling, Anal. Chem. 77 (4) (2005) 964–973.

[21] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, et al., BLAST+: architecture and applications, BMC Bioinform. 10 (2009) 421 (Available from: ⟨http://www.ncbi.nlm.nih.gov/pubmed/20003500⟩, ⟨http://download.springer.com/static/pdf/101/art:10.1186/1471-2105-10-421.pdf?⟩, Origin Url=⟨http://bmcbioinformatics.biomedcentral.com/article/10.1186/1471-2105-10-421⟩).