

Dataset brief**Comprehensive proteomic analysis of *Penicillium verrucosum***

Katharina Nöbauer^a, Karin Hummel^a, Corina Mayrhofer^{b,c}, Maike Ahrens^d, Francis M.C. Setyabudi^a, Martin Eisenacher^d and Ebrahim Razzazi-Fazeli^{a*}

^a *VetCore Facility for Research, University of Veterinary Medicine Vienna, Austria*

^b *Institute of Animal Breeding and Genetics, Department for Biomedical Sciences, University of Veterinary Medicine Vienna, Austria*

^c *Institute of Biotechnology in Animal Production, Interuniversity Department for Agrobiotechnology (IFA Tulln), University of Natural Resources and Applied Life Sciences Vienna, Tulln, Austria*

^d *Medizinisches Proteom-Center, Ruhr-University Bochum, Bochum, Germany*

^{*} *corresponding author:*

Ebrahim Razzazi-Fazeli

University of Veterinary Medicine Vienna

VetCore Facility for Research

Veterinärplatz 1, 1210 Vienna

e-mail: Ebrahim.razzazi@vetmeduni.ac.at

Abbreviation

OTA, ochratoxin A; SWATH, Sequential Window Acquisition of all Theoretical fragment-ion spectra; FA, formic acid; PP, ProteinPilot software

Keywords

Received: 30/11/2016; Revised: 23/02/2017; Accepted: 01/03/2017

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the [Version of Record](#). Please cite this article as [doi: 10.1002/pmic.201600467](#).

This article is protected by copyright. All rights reserved.

Fungal proteomics, *Penicillium verrucosum*, shotgun proteomics, SWATH

Total number of words: 2398

Abstract

Mass spectrometric identification of proteins in species lacking validated sequence information is a major problem in veterinary science. In the present study we used ochratoxin A (OTA) producing *Penicillium verrucosum* to identify and quantitatively analyse proteins of an organism with yet no protein information available. The work presented here aimed to provide a comprehensive protein identification of *P. verrucosum* using shotgun proteomics. We were able to identify 3631 proteins in an “ab initio” translated database from DNA sequences of *P. verrucosum*. Additionally a Sequential Window Acquisition of all Theoretical fragment-ion spectra analysis (SWATH) was done to find differentially regulated proteins at two different time points of the growth curve. We compared the proteins at the beginning (day 3) and at the end of the log phase (day 12).

Penicillium verrucosum is a filamentous fungus producing ochratoxin A (OTA) and citrinin. Exposure to these mycotoxins represents a serious health issue and has been associated with several animal and human diseases [1]. To date only 419 reviewed protein entries in the UniProt database (11/16) with the taxonomy “*Penicillium*” and 4288 proteins of the taxonomy “*Aspergillus*” were available. So far no shotgun analyses of *P. verrucosum* have been published. Using 2D-gel electrophoresis and subsequent MS analysis 32 different proteins in 46 spots were identified via homology to other fungal species [2].

Penicillium verrucosum strain BFE808 (kindly provided by Markus Schmidt-Heydt) was pre-cultivated on malt extract agar plates. The spores were stored in 14% sucrose, 1% peptone at -80°C. For cultivation 1.35×10^5 spores were added to 300 ml broth of either malt extract broth or yeast extract broth. Biomass was cultivated in the dark in both media and at constant light exposure just in yeast extracts media. All cultures grew at 22°C. For each experiment two biological replicates of the biomass

were harvested after 3 and 12 days. All 12 samples were used for identification in order to generate comprehensive identifications and subsequent ion libraries.

The mycelium was harvested, washed with PBS, lyophilized and weighed. The dried biomass was ground to a fine powder in liquid nitrogen using a mortar and pestle. Fifty milligram of the powder was dissolved in 700 µl of lysis buffer containing 7 M urea, 2 M thiourea, 4% CHAPS, 1% DTT and protease inhibitor mix (Sigma, USA). After successful protein quantification (Pierce 660nm assay, Thermo Fisher Scientific, USA) 10 µg of protein was digested with a combination of LysC and Trypsin (Promega, USA) according to the manufacturer's information. Separation of 250 ng of peptides was accomplished on a nano-HPLC system (Dionex, The Netherlands) equipped with a 50 cm C18 acclaim pepmap column (Dionex, The Netherlands). For desalting and pre-concentration the peptides were loaded to a trap column using 2% ACN in HPLC water with 0.05% TFA. Then separation was carried out starting with 98% of mobile phase A (0.1% aqueous FA) and 2% of mobile phase B (80% ACN in HPLC water with 0.08% FA) increasing B to 35% in 130 min, followed by a washing step of 10 min. A 5600+ TripleTOF (Sciex, USA) was used for identification (IDA) and quantification (SWATH) runs. For the IDA experiments a 140 ms survey scan (TOF-MS) was performed, from which the top 20 ions were selected for MS/MS analysis, each a 100 ms scan. [3,4]

In a first approach databases of two genetically closely related *Penicillium* species (*P. chrysogenum* and *P. nordicum*) were used for homology-based identification. Database searches were done with following parameters: trypsin as enzyme, precursor tolerance of 10 ppm, MS/MS mass tolerance of 0.05 Da, FDR 1% on protein-, peptide- and spectra-level. Considering proteins with at least 2 peptides per protein using the rapid search algorithm led to identification of 2339 (*chrysogenum* database) and 3297 proteins (*nordicum* database). Allowing more modifications and the additional biological modification "amino acid substitution" in ProteinPilot (Sciex, USA) thorough search [5], the number of identified proteins was strongly increased. As shown in Table 1 totally 2611 proteins (+10%) were identified in the *chrysogenum* database and 3436 proteins (+4%) in the *nordicum* database. The proteomics data have been deposited to the ProteomeXchange Consortium (<http://proteomecentral.proteomexchange.org>) via the PRIDE partner repository [6] with the dataset identifier <PXD005459>. Due to the clearly higher number of identified proteins in the *nordicum* database, this database was used as a training database in AUGUSTUS [7] to "ab initio" translate the available DNA data from *P. verrucosum* [8]. The full genome sequence of isolate BFE 808, comprising 4207 scaffolds or contigs with a total length of

31.1507Mb was translated into 9232 predicted proteins (fasta file available at ProteomeXchange <PXD005459>). To validate the parameters used by AUGUSTUS, the genome of *P. nordicum* was translated into predicted proteins with the same parameters. This translation resulted in 8125 predicted proteins of which 97.5% showed a high homology (>70%) to the *P. nordicum* proteins published in the UniProt.

To get a brief overview on the biological function of the proteins, a PFAM analysis was performed (<http://pfam.xfam.org/>, Supporting Information Table 1). Proteins from 306 different PFAM clans were found, 1490 proteins (16%) could not be associated with any clan. The biggest group was CL0063 (FAD/NAD(P)-binding Rossmann-fold Superfamily). These redox proteins consist of two domains, a catalytic domain and the Rossmann-fold domain respectively. The proteins of the second big group CL0020 (Tetratrico peptide repeat superfamily) are involved in cell cycle regulation, transcriptional control, mitochondrial and peroxisomal protein transport, neurogenesis and protein folding. The third big group CL0023 (P-loop containing nucleoside triphosphate hydrolase superfamily) represents proteins which perform chaperone-like functions and assist in the assembly, operation, or disassembly of protein complexes. The first 15 clans and their number of associated identified proteins are shown in Figure 1.

In a next step the “ab initio” database was used for the identification resulting in 3631 identified proteins. The number of identified spectra was increased to 86.4% of all spectra searched (Table 1). Even not allowing amino acid substitutions (ProteinPilot rapid search) 3488 proteins are identified (Venn diagram in Figure 2) showing these sequences exactly fit to the measured MS data. All search results are available at ProteomeXchange <PXD005459> as well as in Supporting Information Table 2. To compare the translated proteins of *P. verrucosum* to the nordicum database, the translated sequence fasta file of *P. verrucosum* was blasted against the nordicum database using blast 2.4.0+ (NCBI) (Supporting Information Table 3) resulting in 8827 homologous proteins with a percentage of identical matches >70%.

For generating the ion library, the combined search of all samples in the “ab initio” *P. verrucosum* database was used. Quantitative analyses were focused on samples grown on yeast extract in the dark. Further data handling included PeakView SWATH Processing Micro App (Sciex, USA) to find the correct peak group in the set of chromatograms. Import of proteins from PP group files to the ion

libraries was restricted to all proteins, which met the criteria of a global FDR of 1%. After RT alignment, processing parameters were set to: max 6 peptides and 6 transitions, 96% confidence level, 5% FDR, shared peptides excluded for SWATH analysis, 8 min window, 100 ppm width and modified peptides included. 3446 of the identified proteins matched these criteria and were quantified with SWATH (Supporting Information Table 4).

For SWATH MS/MS analysis 35 overlapping windows (20 Da each, 1 Da overlap) over the full mass range of 400 – 1100 m/z were constructed and measured. A cycle time of 3 sec led to an accumulation time of 82 ms for each window [9,10]. After processing raw data, areas were exported to MarkerView 1.2.1 (Sciex, USA) to visualise the data structure based on Principle Component Analysis (PCA) and to normalize peak areas according to total area sums. For determination of differentially regulated proteins, the data was further analysed with R programming language (version 3.2.2) [11]. On the logarithmic scale, technical replicates were aggregated by the arithmetic mean before significant differences between the two time points (3 and 12 days) were examined with unpaired two-sample *t*-test (unequal variances). All resulting *p*-values were adjusted for multiple testing controlling the FDR [12]. Candidates of interest were chosen based on a *p*-value lower than the significance level of $\alpha=0.05$ and an up- or down-regulation of at least 2-fold. Only proteins with at least two peptides used for quantitation were included.

PCA confirmed high reproducibility of biological and technical replicates by grouping the replicates of each time point closely together but separating them according to the two time points. The scores for PC1 and PC2 were 66.8% and 7.2%, respectively, showing that a major part of the overall variation in the data is captured in the first two principal components (Supporting Information Figure 1). As shown in Figure 3a four different proteins were found to be regulated with $p \leq 0.05$. Figure 3b shows the 243 additionally regulated proteins within an adjusted *p*-value of 0.05 and 0.1. A detailed table can be found in Supporting Information Table 4 and 5.

Looking at the biological functions of the regulated proteins the majority of ones down-regulated at the end of the log-phase of the growth experiment is involved in protein biosynthesis and ribosome biogenesis. These processes seem to be reduced in the fungus at day 12 due to limited amounts of nutrients in the media. The up-regulated proteins on the other hand are associated with proteolysis, amino acid metabolic processes, gluconeogenesis and carbohydrate metabolic processes compensating the low amount of nutrients left.

In summary this study is the first shotgun proteomics analysis of *P. verrucosum* evaluating an “ab initio” gene prediction database identifying up to 3631 proteins of this filamentous fungus. This expanded the gel based approach by Stoll et al including the identification of 31 out of the 32 different identified proteins. SWATH analysis showed that the fungus reduced ribosomal biogenesis and protein biosynthesis in the nutrient limited environment and enhanced protein and carbohydrate processing. The use of the “ab initio” translated database to identify proteins of this yet not annotated organism serves as a profound basis for further investigations on the influence of growth conditions on *P. verrucosum*.

Acknowledgment

We thank Anja Friedrich for her excellent contribution during manuscript preparation.

The statistical part of the study was supported by de.NBI (FKZ 031 A 534 A), a project of the BMBF (Bundesministerium für Bildung und Forschung) and by PURE (Protein research Unit Ruhr within Europe) funded by the Ministry of Science, North Rhine-Westphalia, Germany.

The authors have declared no conflict of interest

References

- [1] Edite Bezerra da Rocha, M., Freire, F. da C.O., Erlan Feitosa Maia, F., Izabel Florindo Guedes, M., Rondina, D., Mycotoxins and their effects on human and animal health. *Food Control* 2014, 36, 159–165.
- [2] Stoll, D. a., Link, S., Kulling, S., Geisen, R., Schmidt-Heydt, M., Comparative proteome analysis of *Penicillium verrucosum* grown under light of short wavelength shows an induction of stress-related proteins associated with modified mycotoxin biosynthesis. *Int. J. Food Microbiol.* 2014, 175, 20–29.
- [3] Rizwan, M., Miller, I., Tasneem, F., Böhm, J., et al., Proteome analysis of *Aspergillus ochraceus*. *Mycotoxin Res.* 2010, 26, 171–180.
- [4] Kumar, G., Hummel, K., Ahrens, M., Menanteau-Ledouble, S., et al., Shotgun proteomic analysis of *Yersinia ruckeri* strains under normal and iron-limited conditions. *Vet. Res.* 2016, 47, 1–13.
- [5] Shilov, I. V, Seymour, S.L., Patel, A. a, Loboda, A., et al., The Paragon Algorithm, a next generation

search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra. *Mol. Cell. Proteomics* 2007, 6, 1638–1655.

- [6] Vizcaino, J.A., Côté, R.G., Csordas, A., Dienes, J.A., et al., The Proteomics Identifications (PRIDE) database and associated tools: Status in 2013. *Nucleic Acids Res.* 2013, 41, 1063–1069.
- [7] Hoff, K.J., Stanke, M., WebAUGUSTUS--a web service for training AUGUSTUS and predicting genes in eukaryotes. *Nucleic Acids Res.* 2013, 41, 1–6.
- [8] Stoll, D.A., Geisen, R. and Schmidt-Heydt, M., Whole Genome Shotgun Sequence of the Ascomycete *Penicillium verrucosum*. *unpublished* 2015.
- [9] Huang, Q., Yang, L., Luo, J., Guo, L., et al., SWATH enables precise label-free quantification on proteome scale. *Proteomics* 2015, 15, 1215–1223.
- [10] Grobber, C., Viridis, B., Nouwens, A., Harnisch, F., et al., Use of SWATH mass spectrometry for quantitative proteomic investigation of *Shewanella oneidensis* MR-1 biofilms grown on graphite cloth electrodes. *Syst. Appl. Microbiol.* 2014, 38, 135–139.
- [11] R Development Core Team, *R: A language and environment for statistical computing.*, 2009.
- [12] Benjamini, Y., Hochberg, Y., Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* 1995, 57, 289–300.

Figure Legends

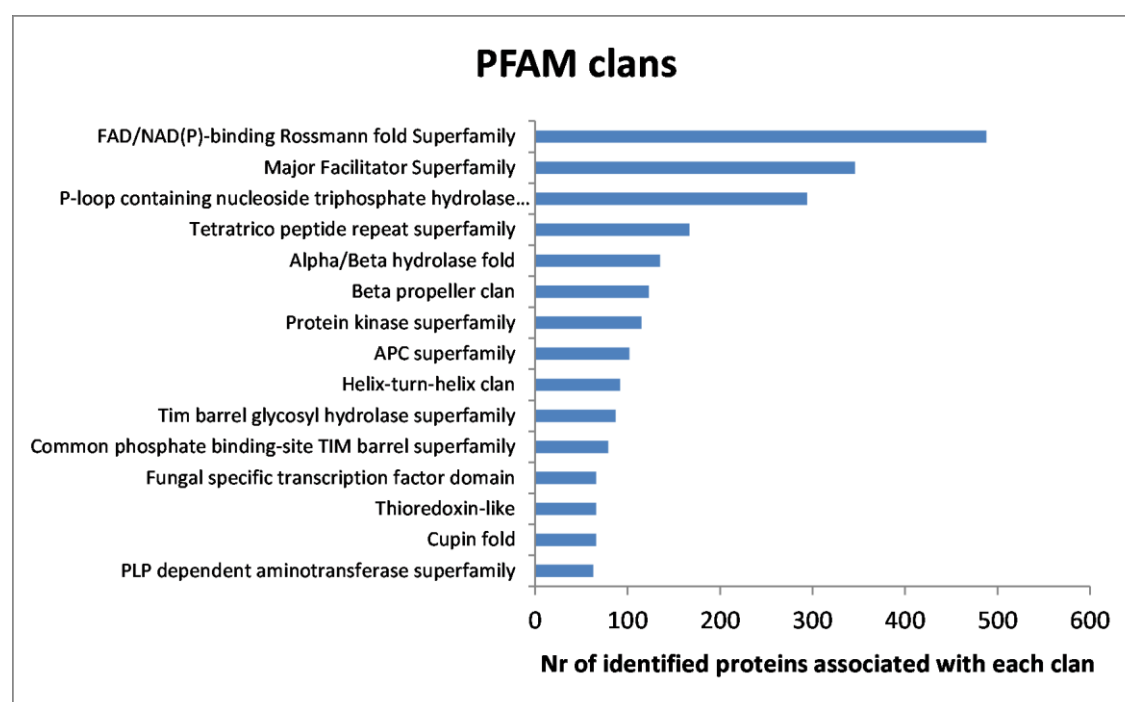


Figure 1: Overview of the 15 PFAM clans with the most associated identified proteins of the "ab initio" translated *P. verrucosum* database.

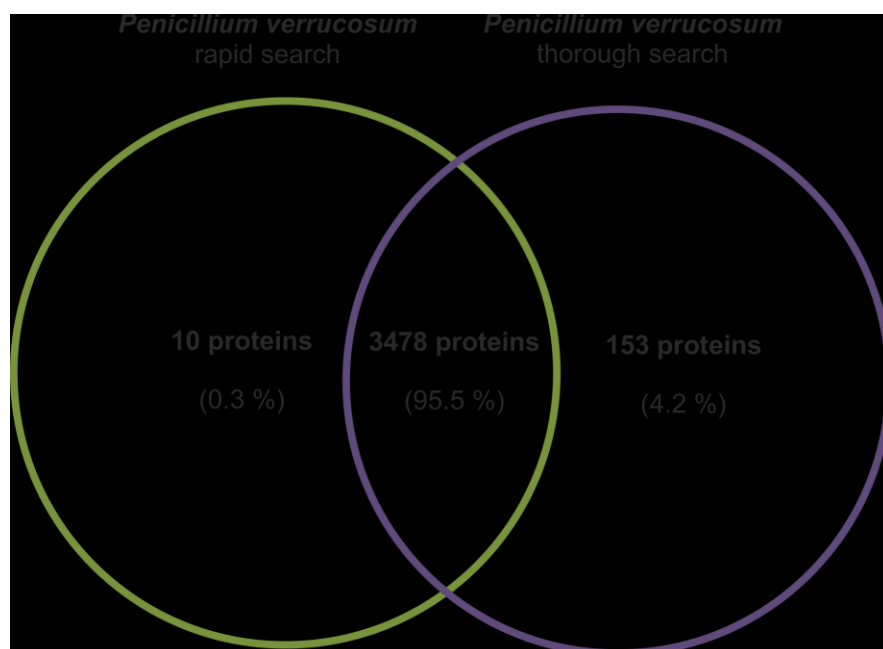


Figure 2: Overlap of identified proteins of the rapid and thorough search in the "ab initio" translated *P. verrucosum* database. Only proteins identified with at least 2 peptides/protein were taken into account.

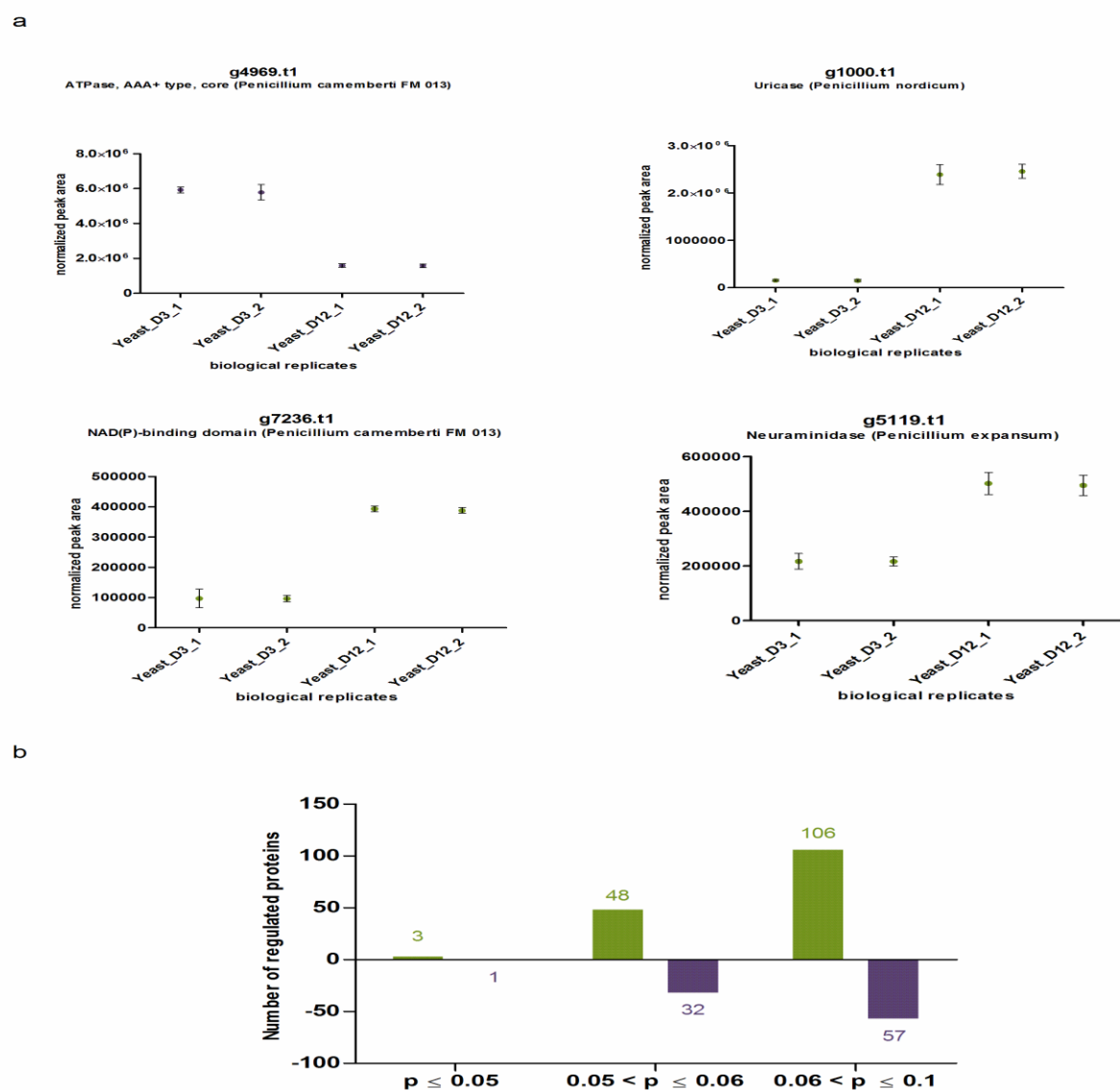


Figure 3: a, Peak areas of significantly regulated proteins below the chosen significance level of 0.05 and with at least 2-fold up- or down-regulation. Values represent the mean of the technical replicates. Error bars show the standard deviation.

b, At least two fold regulated proteins with corresponding adjusted p-values

Tables

Table 1: Comparison of the three different databases used. The table shows the percentage of identified proteins in each database and the percentage of identified spectra searched.

	<i>P. chrysogenum</i>	<i>P. nordicum</i>	<i>P. verrucosum</i>
Percentage of identified proteins per database	20.4% (2611 of 12805 proteins in the DB)	26.5% (3436 of 12983 proteins in the DB)	39.3% (3631 of 9232 proteins in the DB)
Identified spectra (total number of spectra searched: 695,169)	71.3% (495,921)	84.7% (588,940)	86.4% (600,455)