# Combination of Proteogenomics with Peptide *De Novo* Sequencing Identifies New Genes and Hidden Posttranscriptional Modifications

**B. Blank-Landeshammer,[a] I. Teichert,[b] R. Märker,[b] M. Nowrousian,[b,c] U. Kück,[b] A. Sickmann[a]**

[a]Leibniz-Institut für Analytische Wissenschaften-ISAS-e.V., Dortmund, Germany
[b]Allgemeine und Molekulare Botanik, Ruhr-Universität, Bochum, Germany
[c]Lehrstuhl für Molekulare und Zelluläre Botanik, Ruhr-Universität, Bochum, Germany

**ABSTRACT** Proteogenomics combines proteomics, genomics, and transcriptomics and has considerably improved genome annotation in poorly investigated phylogenetic groups for which homology information is lacking. Furthermore, it can be advantageous when reinvestigating well-annotated genomes. Here, we applied an advanced proteogenomics approach, combining standard proteogenomics with peptide *de novo* sequencing, to refine annotation of the well-studied model fungus *Sordaria macrospora*. We investigated samples from different developmental and physiological conditions, resulting in the detection of 104 so-far hidden proteins and annotation changes in 575 genes, including 389 splice site refinements. Significantly, our approach provides peptide-level evidence for 113 single-amino-acid variations and 15 C-terminal protein elongations originating from A-to-I RNA editing, a phenomenon recently detected in fungi. Coexpression and phylostratigraphic analysis of the refined proteome suggest that new functions in evolutionarily young genes correlate with distinct developmental stages. In conclusion, our advanced proteogenomics approach supports and promotes functional studies of fungal model systems.

**IMPORTANCE** Next-generation sequencing techniques have considerably increased the number of completely sequenced eukaryotic genomes. These genomes are mostly automatically annotated, and *ab initio* gene prediction is commonly combined with homology-based search approaches and often supported by transcriptomic data. The latter in particular improve the prediction of intron splice sites and untranslated regions. However, correct prediction of translation initiation sites (TIS), alternative splice junctions, and protein-coding potential remains challenging. Here, we present an advanced proteogenomics approach, namely, the combination of proteogenomics and *de novo* peptide sequencing analysis, in conjunction with Blast2GO and phylostratigraphy. Using the model fungus *Sordaria macrospora* as an example, we provide a comprehensive view of the proteome that not only increases the functional understanding of this multicellular organism at different developmental stages but also immensely enhances the genome annotation quality.

**KEYWORDS** proteogenomics, peptide *de novo* sequencing, RNA editing, alternative splicing, phylostratigraphy, gene ontology, alternative splice sites, fungal genome, genomics, proteomics

Mass spectrometry (MS)-based proteomics has been established as a valuable tool for detecting and quantifying peptides and posttranslational modifications (PTMs) on a large scale. Here, experimental tandem MS (MS/MS) spectra are compared with theoretical spectra derived from curated protein sequence databases. However,

**Citation** Blank-Landeshammer B, Teichert I, Märker R, Nowrousian M, Kück U, Sickmann A. 2019. Combination of proteogenomics with peptide *de novo* sequencing identifies new genes and hidden posttranscriptional modifications. mBio 10:e02367-19. https://doi.org/10.1128/mBio.02367-19.

**Editor** Joseph Heitman, Duke University

Address correspondence to U. Kück, ulrich.kueck@rub.de, or A. Sickmann, albert.sickmann@isas.de.

This article is a direct contribution from U. Kück, a Fellow of the American Academy of Microbiology, who arranged for and secured reviews by Oliver Valerius, Georg-August University, and Andrea Urbani, Catholic University of the "Sacred Heart," Rome.

**Received** 11 September 2019
**Accepted** 13 September 2019
**Published** 15 October 2019

novel findings are inherently limited by the database provided. Thus, expansion of this search strategy to the use of 6-frame translations of the reference genome, theoretically predicted protein sequences, and transcriptome sequencing (RNA-Seq)-derived transcript sequences gave rise to so-called proteogenomics approaches (1, 2). This enabled the reannotation of genomes, the correction of misannotated genes, the discovery of novel protein-coding regions, and the detection of alternative translation initiation and termination sites. Proteogenomics refinements are now being routinely applied to prokaryotic genomes, and several examples are available for plant, animal, and human experimental systems (2–6). However, particularly in the case of fungal proteogenomics, investigated samples were often limited to only some developmental stages (7–10).

Although providing peptide-level evidence for coding sequences and, thus, being of high value for genome (re-)annotation, proteogenomics analysis does not identify peptides beyond the DNA-based coding potential, i.e., peptides modified by posttranscriptional changes at the RNA level. Here, we present an advanced two-enzyme proteogenomic workflow, which was extended by *de novo* peptide sequencing, curation, and validation. We applied this advanced proteogenomics workflow to the well-annotated genome sequence of the fungal model *Sordaria macrospora* (11, 12), using samples from five different developmental or physiological conditions. *S. macrospora* is used as a model organism for multicellular development during the fungal sexual cycle (13, 14) and has been scrutinized for signaling pathways and conserved developmental regulators governing sexual fruit-body formation (15, 16). However, in-depth proteomics analysis has not been performed with *S. macrospora*, and information about the stage-specific proteome is lacking.

Applying the advanced proteogenomics approach to five *S. macrospora* samples not only increased the number of annotated protein-coding genes but also provided evidence for alternatively spliced gene variants, extensions, fissions, and fusions. Most importantly, the combination with peptide *de novo* sequencing led to the discovery of so-far undetectable peptides originating from A-to-I editing at the mRNA level. Our approach provides quantitative data about stage-specific protein abundances, leading to new predictions for the biological role of certain proteins at distinct developmental stages, and will enable further applications, such as evolutionary genetic investigations.

## RESULTS

**Annotation refinements by proteogenomics.** We previously sequenced the genome of the *S. macrospora* wild-type strain (termed genome version 1 [v1]) and later refined the genome annotation by implementing RNA-Seq data (genome v2) (11, 12). Regeneration of the sequenced strain by several sexual crosses led to ascospore progeny showing vigorous growth and robust fruiting body development. We chose ascospore isolate R19027 for Illumina sequencing and performed annotation based on genome v2 and RNA-Seq data (12, 17, 18). Locus tags were transferred from genome v2, followed by manual refinement of about 1,000 genes (genome v3). However, to add peptide-level evidence for annotated features and to elucidate hidden coding potential, we also applied a combination of proteogenomics with peptide *de novo* sequencing.

A prerequisite for any proteogenomics study is deep coverage of the target proteome. To trigger divergent gene expression, *Sordaria macrospora* was grown for 3 days on three culture media with different nutrient compositions (Biomalz-Mais-Medium [BMM], complete medium [CM], and Sordaria Westergaard's medium [SWG]). Furthermore, BMM cultures were harvested at three time points (2 days of shaking culture and 3 and 7 days of surface culture), leading to five different conditions in total. While only vegetative cells were present in 2-day shaking cultures, 3- and 7-day surface cultures contained immature and subsequently mature fruiting bodies. To achieve high coverage of proteins and identify splice site-specific peptides, samples were digested with two endoproteinases of complementary specificity, i.e., trypsin and Glu-C, and were subjected to fractionation by high-pH reversed-phased chromatography prior to liquid chromatography MS/MS (LC-MS/MS) analysis. The total data set comprised 4,027

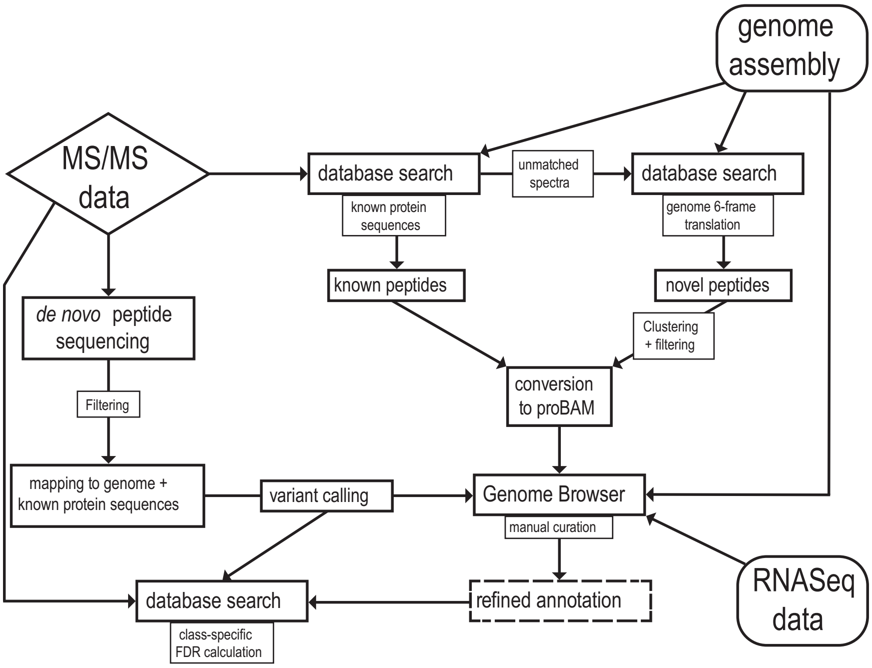


**FIG 1** Schematic representation of the data analysis pipeline. Generated MS/MS spectra were subjected to subsequent database searches against the known *S. macrospora* protein sequences and a 6-frame translation of the *S. macrospora* genome as well as an independent *de novo* peptide sequencing method. Putative novel peptide identifications were clustered, filtered, converted to a genome browser readable format, and analyzed in conjunction with RNA-Seq data. Final curation of the genome annotation was performed manually.

million MS/MS spectra acquired with high resolution from a total of 128 LC-MS runs (see Table S1 in the supplemental material).

Spectra were analyzed via a multistep proteogenomics analysis workflow (Fig. 1) and subjected to database searches against an *S. macrospora* protein database generated from genome v3, a 6-frame genome translation, and unbiased *de novo* peptide sequencing, followed by thorough filtering and manual annotation refinement in conjunction with RNA-Seq data (12). In total, 6,223 proteins were identified with a 1% false discovery rate threshold (FDR) (protein level), representing 62% of the total *S. macrospora* proteome, reaching an average sequence coverage of 45%. Four hundred ten proteins were exclusively identified in the 7-day sample, representing the growth condition with the largest contribution to total proteome coverage. Translation initiation sites (TIS) of 2,006 proteins were validated by identifying N-terminally acetylated peptides. A total of 13,075 peptides were identified as exon spanning, providing peptide-level evidence for 4,723 splice junctions, 659 of which were exclusively identified by Glu-C-derived peptides. Overall, 45% of all possible splice junctions within the identified proteome were covered.

Comparing the complete data set with *S. macrospora* genome v3, we identified 7,803 novel peptides, representing 4% of all detected peptides. While class-specific FDR (known and novel peptides) was accounted for in the refinement process by the stepped search strategy against predicted proteins and the genome, we also wanted to assess the quality of the identifications after the final refinement. Comparison of peptide class-specific identifications with known false-positive decoy hits above the threshold showed no significant difference between the classes with regard to average peptide length, mass deviation, or posterior error probability (PEP) (Fig. 2A to C and Fig. S1 and S2A). We further compared the chromatographic retention behavior of the novel identifications with the predicted peptide hydrophobicity index (HI). Observed chromatographic retention times strongly correlated with SSRCalc-predicted HI values

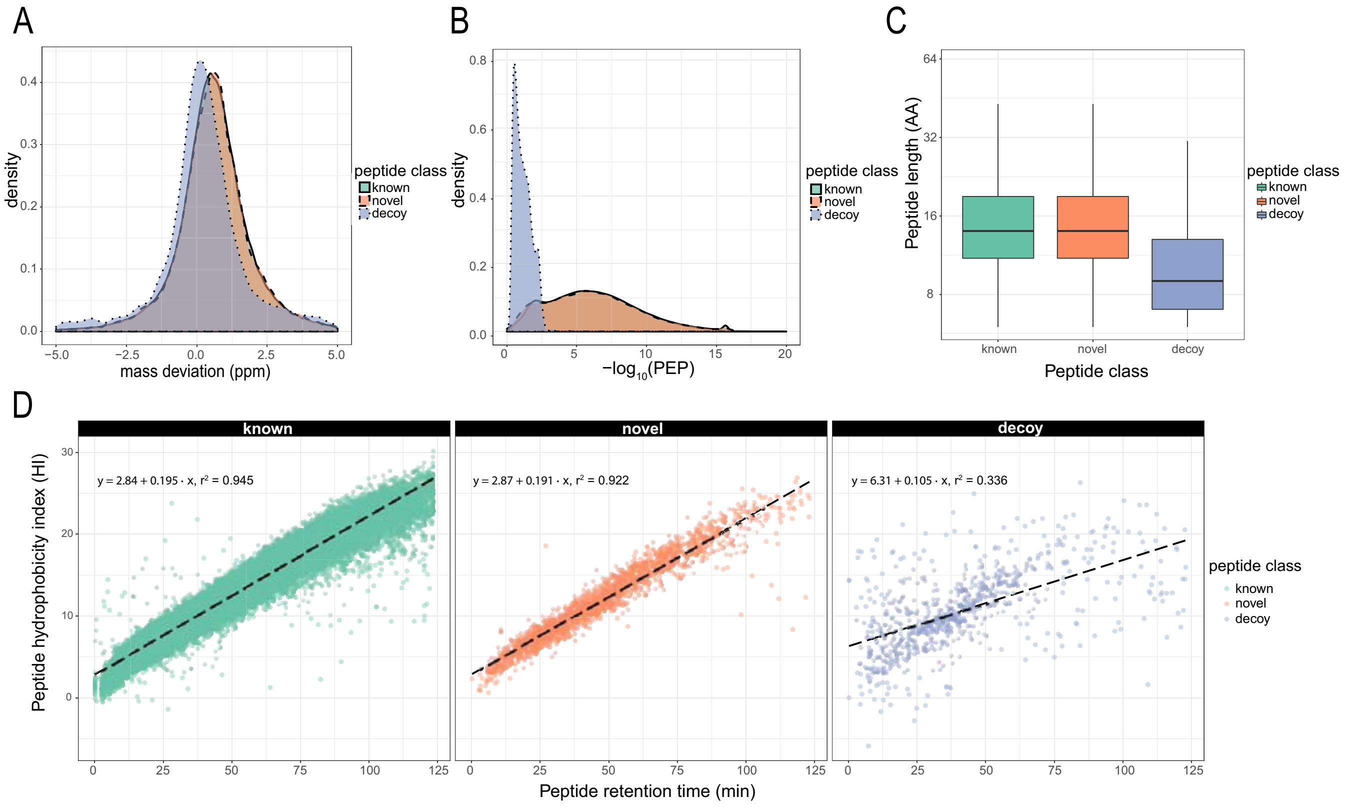


**FIG 2** Evaluation of peptide identifications, classified as known and novel, in the 2-day data set. All peptide spectrum matches (PSMs) belonging to the respective class were compared to known false-positive decoy hits. (A) Normalized density plot of observed precursor mass deviation indicates no difference in distribution of identified known and novel PSMs. (B) Distribution of posterior error probability (PEP) shows clear distinction between decoy PSMs and known PSMs but almost overlapping distribution of known and novel hits. (C) Length of identified peptides of both classes, but not between classes, differs from decoy identifications. AA, amino acids. (D) Observed retention time shows tight correlation to predicted hydrophobicity index (HI) by SSRCalc for both classes, while retention times of known false positives only weakly correlate. In all cases, all decoy PSMs with a q value of <0.01 were plotted as a reference. See Fig. S1 and S2 for additional data sets.

(19) for both known and novel peptides, whereas known false-positive decoy peptides showed a scattered distribution (Fig. 2D and Fig. S2B).

Based on the proteogenomic refinements, we newly annotated 575 genes, corresponding to 6.7% of all protein-coding genes (Table 1 and Data Set S1). Correction of splice sites was most frequent, accounting for one-third of all refinement operations. A further 25% of refinement operations led to the correction of annotated TIS or the addition of alternative TIS, up- or downstream relative to the annotated TIS.

**Distinct GO terms are enriched in defined growth phases.** Following functional annotation with Blast2GO, gene ontology (GO) enrichment analysis was performed on proteins identified uniquely in a single culture condition to deepen insights into fungal development and metabolism, as well as to cross-validate our results (Fig. S3A and B). While unique proteins identified for *S. macrospora* grown on BMM, SWG, or CM medium for 3 days did not reveal statistically significantly enriched terms, the 410 proteins identified uniquely in the 7-day BMM sample uncovered 38 significantly enriched GO terms (adjusted *P* value of <0.05). The cellular compartment ontology was populated mainly by terms related to vacuole, endoplasmic reticulum (ER), and membrane functions (Fig. S3C). Accordingly, metabolic process terms related to cell wall organization and transmembrane transport were enriched along with those for secondary metabolites and steroid biosynthesis. In contrast, the 115 unique identifications in the 2-day BMM sample overrepresented proteins related to sphingolipid biosynthesis (Fig. S3D).

**Discovering hidden proteins.** Besides annotation modifications of known genes, our data provide evidence for 104 hidden protein-coding genes not accounted for in genome v3 (Table 1). We distinguished two classes. Class I proteins were identified by

**TABLE 1** Classification of all annotation refinements performed in this study

| Type of refinement | Total no. of refinements | No. of refinements with two variants |
|---|---|---|
| Splice site (3′ or 5′ splicing site) | 389 | 45 |
| TIS | 283 | 101 |
| Annotation extension (up- or downstream of annotation) | 237 | |
| Frame shift | 116 | |
| Annotation fission | 13 | |
| Annotation fusion | 7 | |
| Novel annotation | 104 | |

2 or more unique, unmodified peptides longer than 8 amino acids. Furthermore, annotation was based on sufficient RNA-Seq coverage. For class II proteins, only one unmodified unique peptide was found, and their genes had insufficient RNA-Seq coverage. Ninety genes were categorized as class I, while 14 were categorized as class II. Twenty-three of those hidden genes are located at contig borders and, thus, were only partially annotated, with translation sequences lacking the mature protein C and/or N terminus. For 27 novel proteins, annotated TIS were verified by detecting N-terminally acetylated peptides. Remarkably, 18 of the hidden proteins were exclusively identified in the 7-day sample, i.e., the final stage of the reproduction cycle of *S. macrospora* when the fungus produces mature fruiting bodies. For those for which we could identify homologous proteins, functional associations to cell wall and cytoskeleton organization or stress responses were observed. Using Blast2GO analysis, we assigned one or more GO terms to 65 of those hidden proteins. Among those, proteins showing resemblance to *S*-adenosyl-L-methionine-dependent methyltransferases are clearly overrepresented, with 12 novel annotations, followed by 8 heterokaryon incompatibility (HET-E) protein homologues (Data Set S1).

Although we could categorize the majority of hidden proteins with Blast2GO analyses, the function of some gene products remained elusive. Quantification at multiple growth stages facilitates functional coexpression analysis by placing these hidden proteins into a quantitative context with well-annotated proteins, potentially allowing for functional connection. Therefore, we conducted one-dimensional LC-MS measurements for label-free quantification. In addition to the samples described above, we added a 5-day BMM sample to better reflect fungal development over time. We quantified a total of 3,592 proteins (1% FDR), comprising 42 (40%) of the newly identified proteins. In 25 cases the highest expression levels were found in either the 2- or 7-day samples, further demonstrating the usefulness of disparate growth conditions in the context of proteogenomics reannotation.

Coexpression analysis also provided a putative function for the newly identified protein SMAC_12925.3, which displayed no homology to known proteins or known protein domains. SMAC_12925.3 was assigned to module 2, which had an expression profile that strongly correlates with fungal sexual development. Subclusters of module 2 were enriched in proteins associated with vacuolar proteolysis, suggesting a putative vacuolar role for SMAC_12925.3. Detailed coexpression analysis methods and results are provided as supplemental material (Text S1 and Fig. S4).

**Uncovering alternative splice sites.** Our proteogenomics data set led to 389 splice site refinements. We detected a total of 45 alternative splicing (AS) events, caused by intron retention (20), alternative 5′ donor sites (5), alternative 3′ acceptor sites (5), mutually exclusive exons (5), and exon skipping (2). For 21 out of 45 AS events, we identified uniquely matching peptides for both protein isoforms. These results confirm the AS events originally seen in the RNA-Seq analysis and further illustrate that the alternative mRNAs are translated into the protein isoforms. BLAST homology searches confirmed that 80% of these events are conserved in other fungal species.

We were able to quantify 17 AS events in five cases in an isoform-specific manner. As illustrated in Fig. 3, AS was observed for the *mek2* transcript, which encodes a

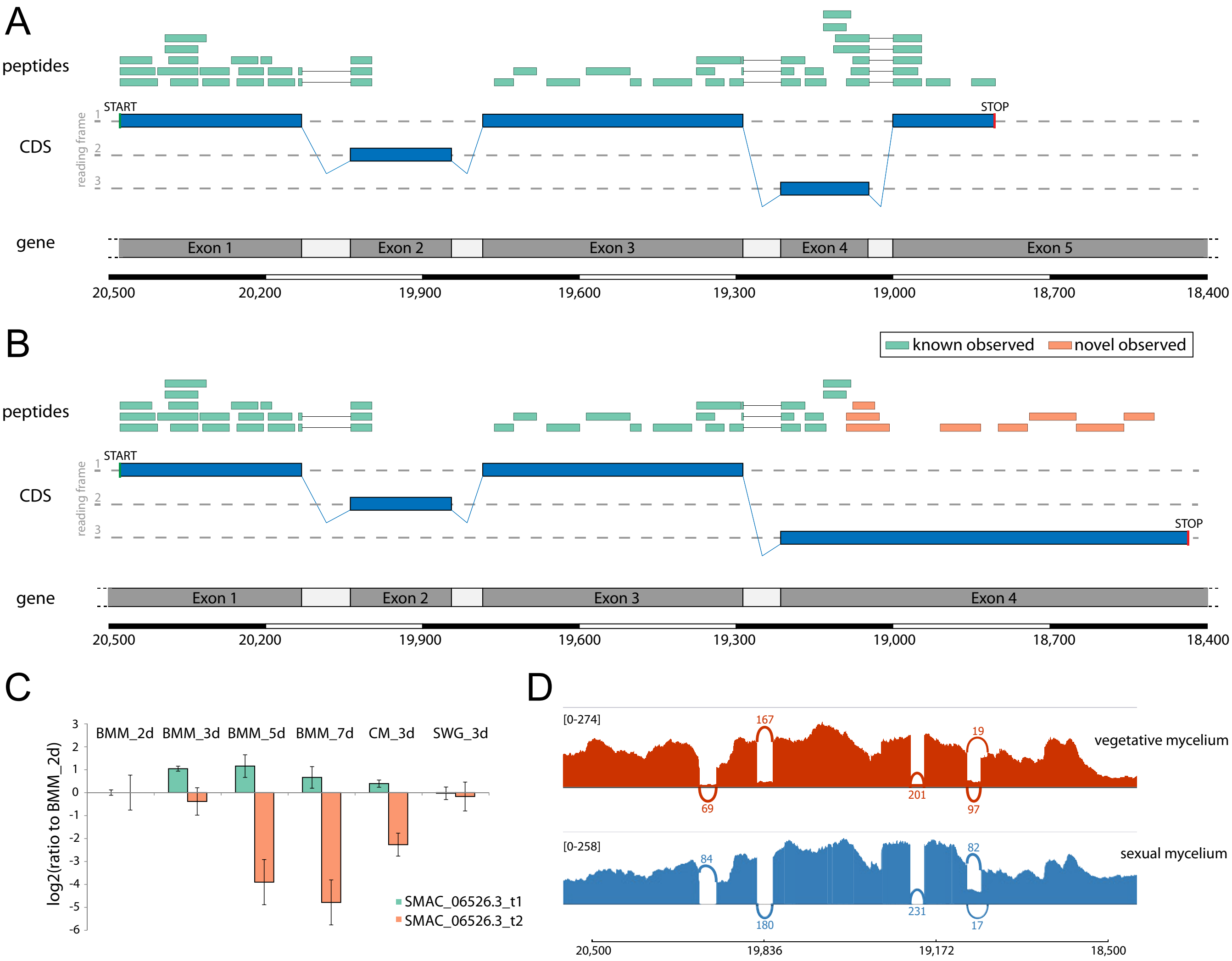


**FIG 3** Alternative splicing in the pheromone pathway-specific kinase gene *mek2*. (A) Graphical representation of the canonical gene structure, including observed peptides (green bars), covering three splice junctions. The *mek2* gene comprises 5 exons and was identified with a total of 50 peptides, covering 75% of the sequence. (B) Graphical representation of the alternatively spliced variant. Retention of intron 4 leads to translation into an alternative protein C terminus, identified by 6 novel peptides (orange bars). (C) Label-free quantification of both MEK2 isoforms throughout six growth conditions of *S. macrospora* reveals downregulation of the newly identified variant (orange, SMAC_06526.3_t2) during sexual development (BMM_5d and BMM_7d). (D) Sashimi plot visualizing the RNA-Seq coverage of both splice variants in vegetative and sexual mycelium.

mitogen-activated kinase (MAPK) with homology to MAPKs of the yeast pheromone signaling pathway (21). The alternatively spliced variant of *mek2* in *S. macrospora* revealed by our proteogenomics search strategy results in translation into an alternative protein C terminus. Label-free quantification further revealed differential regulation patterns of the two variants: while expression of the short variant remained stable under all conditions, the alternatively spliced variant was strongly downregulated at later stages of the fungal reproduction cycle (Fig. 3C). This analysis demonstrates again that the expression of isoforms is linked to specific developmental stages.

**RNA editing.** Beyond using proteogenomics analysis to refine known genes and to identify hidden genes, the data were mined for peptide sequence variations caused by putative changes at the RNA level, i.e., RNA editing. RNA editing is found in all domains of life, and A-to-I mRNA editing has recently been described in filamentous ascomycetes, where editing frequencies have been linked to sexual development (22–25).

To identify A-to-I editing events leading to single-amino-acid variants (SAAVs), we implemented a *de novo* peptide sequencing approach, followed by thorough filtering and refinement steps and additional class-specific database searches with individual

**TABLE 2** Classification of identified SAAVs by type[a]

| SAAV[b] | Mass shift (Da) | No. of observations | Edited and nonedited detected (%) | Median PROVEAN score | Predicted deleterious (%) |
|---|---|---|---|---|---|
| K→E | 0.94763 | 32 | 53 | −1.2565 | 21.9 |
| I→V | −14.01565 | 19 | 95 | −0.689 | 0 |
| S→G | −30.01057 | 14 | 100 | −1.8475 | 28.6 |
| R→G | −99.07965 | 11 | 90 | −3.349 | 63.6 |
| K→R | 28.00615 | 10 | 70 | −1.1875 | 10.0 |
| T→A | −30.01057 | 7 | 100 | −0.792 | 42.9 |
| Q→R | 28.04253 | 6 | 100 | −0.797 | 16.7 |
| Y→C | −60.05414 | 5 | 100 | −6.16 | 80.0 |
| M→V | −31.97208 | 5 | 75 | −1.047 | 20.0 |
| E→G | −72.02113 | 2 | 100 | −2.833 | 50.0 |
| N→S | −27.0109 | 1 | 100 | −2.053 | 0 |
| I→M | 17.95643 | 1 | 100 | −2.867 | 100 |
| Total |  | 113 | 81 | −1.220 | 26.8 |

[a] Total number of observed single-amino-acid variation (SAAV) putatively caused by mRNA editing events in the 7-day sample. Respective theoretical mass shifts of each amino acid exchange are given. Additionally, percentages of sites found for both the edited as well as the nonedited variant are given. PROVEAN prediction score was retrieved for each individual SAAV via the PROVEAN web interface (70), and the median score for each class of SAAV was calculated. A default threshold of −2.5 was used to estimate the extent of potentially deleterious variants.

[b] N→D events were not considered, as they can be caused by RNA editing as well as by asparagine deamidation on the protein/peptide level.

FDR calculation for SAAV peptides. Table 2 shows type and frequency of identified SAAVs and the respective mass shift caused by the altered amino acid for 12 amino-acid substitutions that can be caused by A-to-I editing. We identified a total of 113 SAAVs in 102 distinct proteins. Lysine-to-glutamic acid (K→E) exchange was the most prevalent event, found at 34 different sites, while isoleucine-to-methionine (I→M) exchange was only identified once. We predominantly identified both canonical and edited peptide variants, indicating either substoichiometric or context-specific editing of transcripts. In most cases, a single SAAV per protein was detected, with the exception of putative ubiquitin hydrolase SMAC_02643, containing three SAAVs (K192E, R556G, and S1034G).

To validate these findings, parallel reaction monitoring (PRM) LC-MS runs were conducted to monitor peptide sequence-specific transitions throughout the chromatographic elution profile. Seventeen RNA editing-specific peptides were chosen to represent 10 classes of amino acid substitutions. We then monitored the corresponding nonedited peptide and one proteotypic peptide for every protein, resulting in a total of 51 peptides obtained from fungal protein extracts harvested at four different time points (2 days, 3 days, 5 days, and 7 days). In 16 cases, we detected the edited peptide only in the 5- and 7-day samples, while in the case of SMAC_02363 K43R, the indicative peptide YRPGTVALR was found in all four samples (Fig. S5 and Data Set S2). Figure 4 shows initially identified spectra, transitions, and retention times for edited and nonedited peptides, indicative of SMAC_03693 T255A, caused by RNA editing.

In addition to nonsynonymous recoding events, data were mined for stop-loss editing (UAG to UGG), leading to translation beyond the intrinsic stop codon and potentially generating novel protein C termini. Using *de novo* peptide sequencing and data obtained by searches against genome 6-frame translation, followed by class-specific FDR calculation, we were able to identify peptide sequences indicative of stop-loss editing events in 15 different transcripts. Remarkably, four of these were exclusively identified based on *de novo* peptide sequencing data (Data Set S2). Of the 15 stop-loss editing events, 73% and 20% are conserved in the filamentous fungi *Neurospora crassa* and *Fusarium graminearum* (22, 24, 26), respectively. As a representative example, peptide-level evidence for the stop-loss RNA editing event of the white collar 1 transcript (SMAC_03527.3) reveals a C-terminal 131-amino-acid extension (Fig. 5). First detected in *N. crassa*, white collar 1 is a blue-light photoreceptor involved in circadian clock and developmental processes (27). Domain prediction by CD-Search (20) revealed a putative N-terminal class IIa histone deacetylase domain (E value of 7.3e−04) along with a dimer interface within the novel C terminus. This example indicates diversification of protein functions by mRNA editing in fungi.

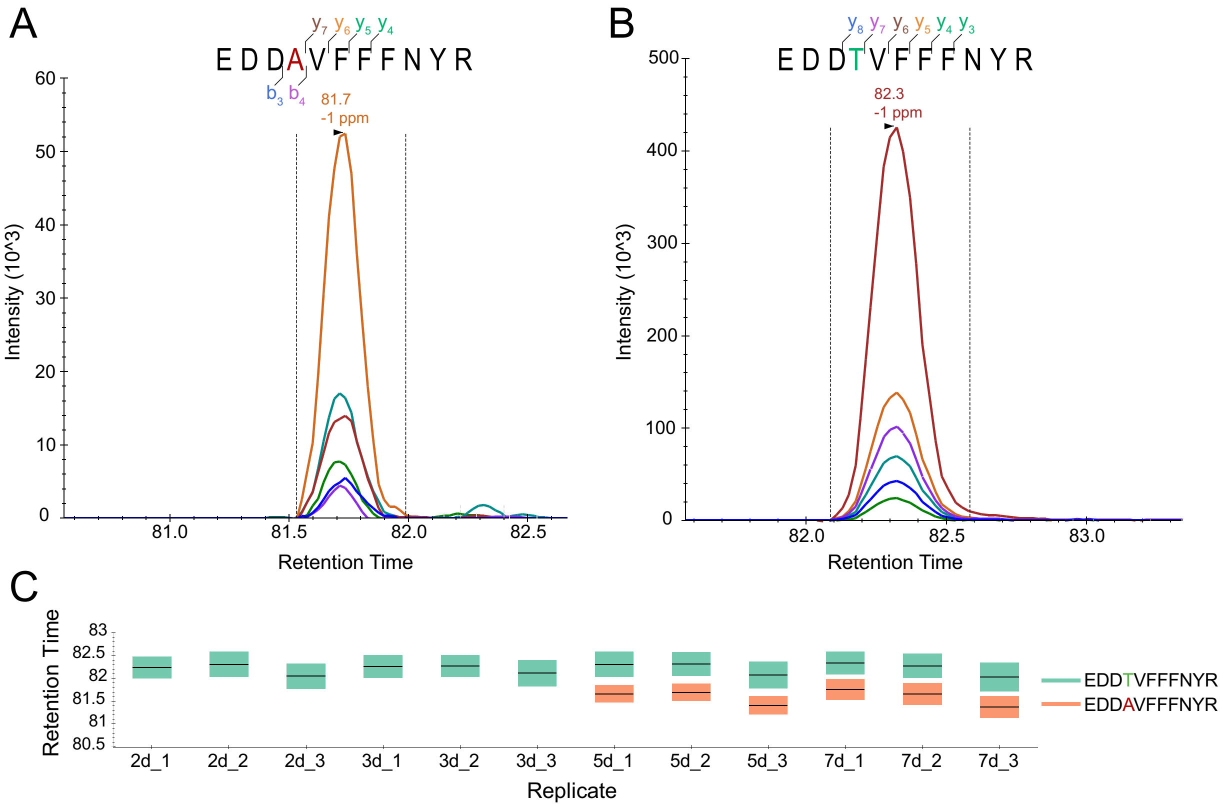


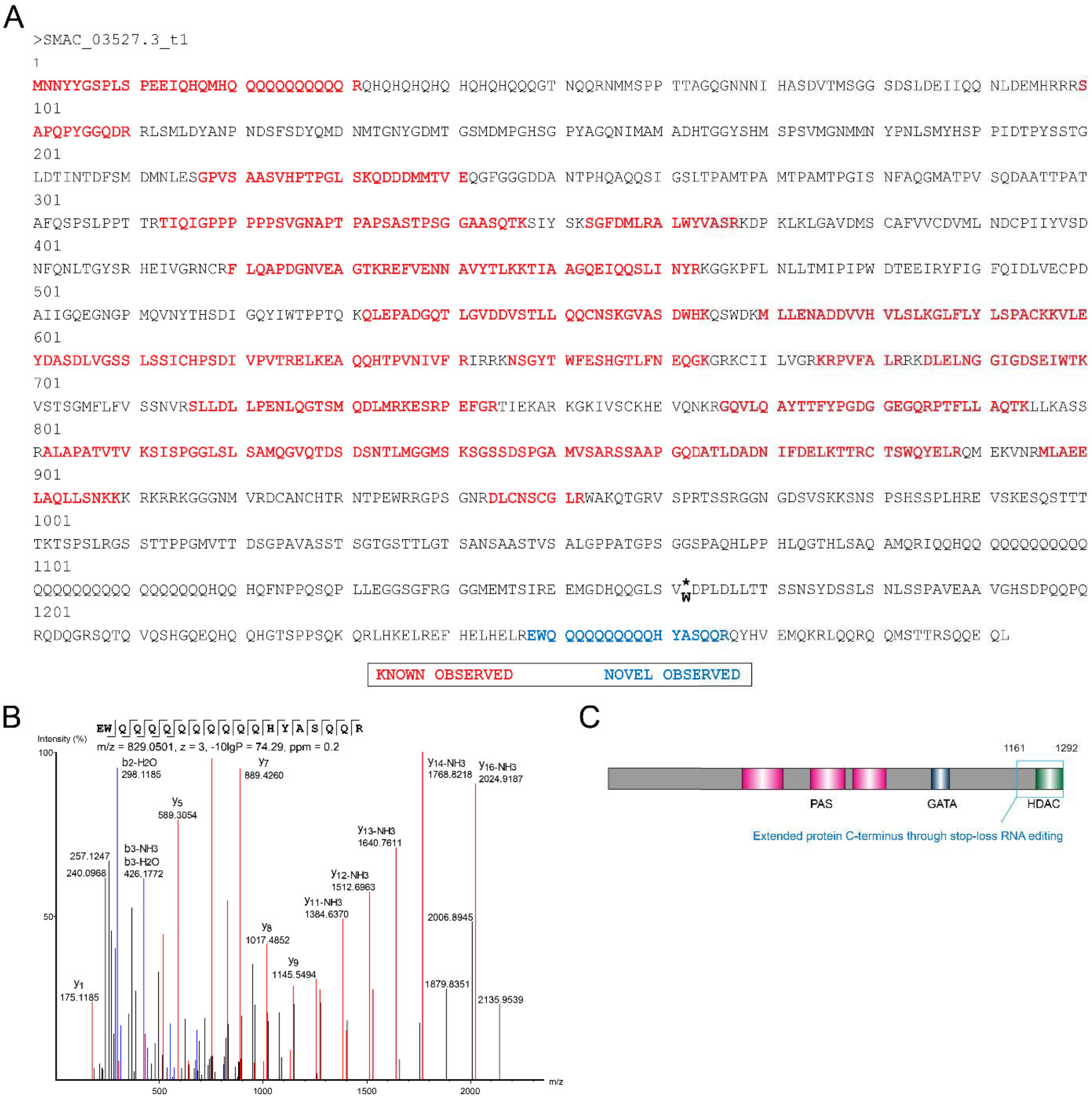
**FIG 4** Peptide-level evidence for a recoding RNA editing event. (A) Parallel reaction monitoring (PRM) transitions of peptide EDDAVFFNYR originating from a putative RNA editing event in SMAC_03693, leading to the exchange of Thr to Ala at position 255. (B) PRM transitions of the genome-encoded peptide EDDTVFFNYR. (C) Peptides were monitored in fungal cells grown for 2, 3, 5, and 7 days, with the peptide originating from edited RNA only being identified in the latter two cases.

**Phylostratigraphy.** Determining the last common ancestor of a given protein can help unravel the emergence of protein families and the evolutionary patterns leading to the origin of genes. Here, we applied phylostratigraphy to determine the evolutionary age of the known and novel proteins of *S. macrospora*. We constructed a phylostratigraphic map and assigned proteins to a total of 15 phylostrata (PS), with PS1 being the evolutionarily oldest and PS15 the evolutionarily youngest phylostratum.

Comparison of the 104 newly identified proteins with all remaining known detected (i.e., identified by proteomics) and known undetected (i.e., not detected by proteomics) proteins of *S. macrospora* shows differences in their relative distribution among the PS (Fig. 6A). Proteins identified by proteomics (i.e., expressed proteins) tended to be more often assigned to an old PS than unidentified proteins (i.e., nonexpressed proteins), which mainly occupy more recent phylostrata. Novel genes showed an intermediate distribution, although they were generally less often assigned to the oldest PS and more often to the comparatively young PS8 to -11 than their known analogues.

The phylostratigraphic map (Fig. 6B) of all fully annotated class 1 hidden proteins gives more detailed insight into their evolutionary emergence. Top BLAST hits are displayed for every PS, and proteins are clustered by Ward's method to facilitate visual differentiation. The topmost clade is mainly occupied by a class of *S*-adenosyl-L-methionine-dependent methyltransferases. These proteins are followed by a clade of highly conserved proteins. These genes mostly have annotated paralogues in the *S. macrospora* genome, likely making them products of gene duplication.

The bottommost clade groups novel proteins of younger evolutionary origin, most of them giving no BLAST hit (below an E value cutoff of 1e−05) beyond PS9, representing functionally specific proteins evolved within the *Sordariomyceta*. Therefore, phylostratigraphic analysis of the refined *S. macrospora* proteome provides evidence for



```
>SMAC_03527.3_t1
1
MNNYYGSPLS PEEIQHQMHQ QQQQQQQQQQ RQHQHQHQHQ HQHQHQQQGT NQQRNMMSPP TTAGQGNNNI HASDVTMSGG SDSLDEIIQQ NLDEMHRRRS
101
APQPYGGQDR RLSMLDYANP NDSFSDYQMD NMTGNYGDMT GSMDMPGHSG PYAGQNIMAM ADHTGGYSHM SPSVMGNMMN YPNLSMYHSP PIDTPYSSTG
201
LDTINTDFSM DMNLESGPVS AASVHPTPGL SKQDDDMMTV EQGFGGGDDA NTPHQAQQSI GSLTPAMTPA MTPAMTPGIS NFAQGMATPV SQDAATTPAT
301
AFQSPSLPPT TRTIQIGPPP PPPSVGNAPT PAPSASTPSG GAASQTKSIY SKSGFDMLRA LWYVASRKDP KLKLGAVDMS CAFVVCDVML NDCPIIYVSD
401
NFQNLTGYSR HEIVGRNCRF LQAPDGNVEA GTKREFVENN AVYTLKKTIA AGQEIQQSLI NYRKGGKPFL NLLTMIPIPW DTEEIRYFIG FQIDLVECPD
501
AIIGQEGNGP MQVNYTHSDI GQYIWTPPTQ KQLEPADGQT LGVDDVSTLL QQCNSKGVAS DWHKQSWDKM LLENADDVVH VLSLKGLFLY LSPACKKVLE
601
YDASDLVGSS LSSICHPSDI VPVTRELKEA QQHTPVNIVF RIRRKNSGYT WFESHGTLFN EQGKGRKCII LVGRKRPVFA LRRKDLELNG GIGDSEIWTK
701
VSTSGMFLFV SSNVRSLLDL LPENLQGTSM QDLMRKESRP EFGRTIEKAR KGKIVSCKHE VQNKRGQVLQ AYTTFYPGDG GEGQRPTFLL AQTKLLKASS
801
RALAPATVTV KSISPGGLSL SAMQGVQTDS DSNTLMGGMS KSGSSDSPGA MVSARSSAAP GQDATLDADN IFDELKTTRC TSWQYELRQM EKVNRMLAEE
901
LAQLLSNKKK RKRRKGGGNM VRDCANCHTR NTPEWRRGPS GNRDLCNSCG LRWAKQTGRV SPRTSSRGGN GDSVSKKSNS PSHSSPLHRE VSKESQSTTT
1001
TKTSPSLRGS STTPPGMVTT DSGPAVASST SGTGSTTLGT SANSAASTVS ALGPPATGPS GGSPAQHLPP HLQGTHLSAQ AMQRIQQHQQ QQQQQQQQQQ
1101
QQQQQQQQQQ QQQQQQQHQQ HQFNPPQSQP LLEGGSGFRG GGMEMTSIRE EMGDHQQGLS V*DPLDLLTT SSNSYDSSLS NLSSPAVEAA VGHSDPQQPQ
                                                                  W
1201
RQDQGRSQTQ VQSHGQEQHQ QHGTSPPSQK QRLHKELREF HELHELREWQ QQQQQQQQQH YASQQRQYHV EMQKRLQQRQ QMSTTRSQQE QL
```

**FIG 5** Validation of a stop-loss editing site in the transcript for the white collar 1 protein. (A) Primary amino acid sequence of the white collar 1 protein. Editing results in the conversion of a stop codon into a tryptophan codon and extends the amino acid sequence by 131 amino acids. As a consequence, the protein carries a C-terminal histone deacetylase domain (HDAC). Peptides encoded by the canonical gene are shown in red, while the blue peptide is unique to the novel C-terminal sequence. (B) Annotated MS/MS spectrum of the editing-specific peptide observed in the 7-day data set. (C) Domain structure of the white collar 1 protein, including the HDAC domain (green box) in the extended C terminus (blue box). PAS, Per-Arnt-Sim domain; GATA, GATA-type zinc finger transcription factor domain.

evolutionarily young genes with specific functions being expressed at distinct developmental stages of fungal life.

## DISCUSSION

The aim of this study was to provide an extensive proteomics data set of *S. macrospora* and use these data in combination with proteogenomics workflows and a novel implementation of peptide *de novo* sequencing to refine the genome annotation

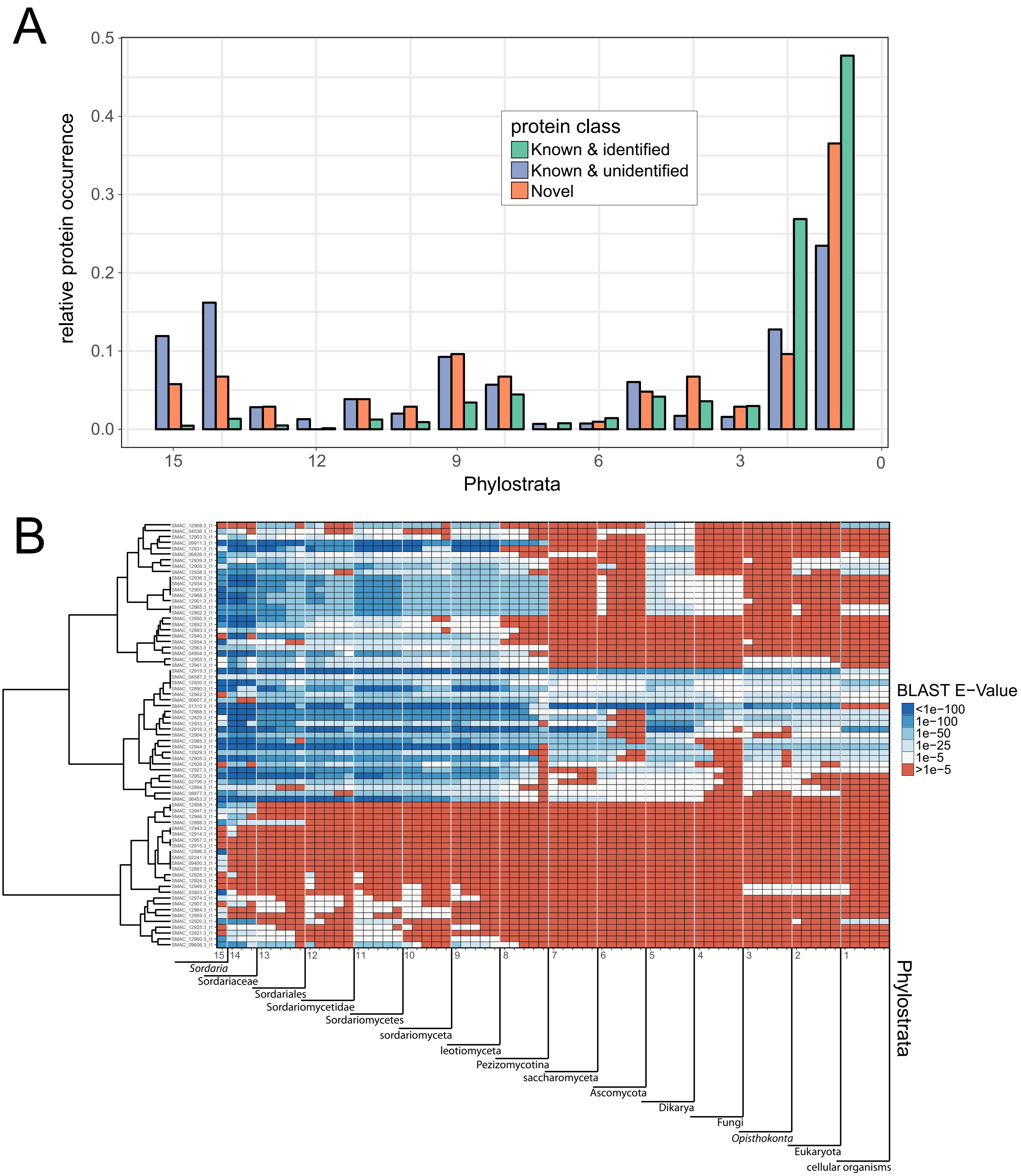


**FIG 6** Phylostratigraphic map of all *S. macrospora* proteins. (A) Division into known and identified detected proteins, novel proteins identified by the proteogenomics analysis, and known and unidentified proteins not found in this study. Relative protein occurrence (i.e., the number of proteins assigned to a PS relative to the total number of proteins) describes the share of proteins assigned to a given phylostratum (PS) within its aforementioned class. (B) Detailed phylostratigraphic map of all novel, completely annotated class I proteins, displaying the BLAST E value of the top 5 hits in every PS. Proteins are hierarchically clustered (Ward's method) to show similarities in PS distribution.

of a fungal model organism and potentially identify posttranscriptional modifications. We further place the whole set of quantitative results within the context of cellular and developmental stage-specific protein abundances.

In total, we were able to confidently identify 6,223 proteins. Four million recorded MS/MS spectra yielded 1,875 million matches to database peptide sequences at a maximum FDR of 1%. Due to the narrow isolation window of 0.4 *m/z* during data acquisition, resulting in exceptionally clear spectra, these data represent an excellent repository of *S. macrospora* peptide spectra, potentially being of use beyond the scope of this study. We were able to identify proteins representing 62% of the theoretical proteome, whereas the remaining 38% was not covered by LC-MS/MS analysis. Arguably, a proportion of these proteins may not be accessible using common proteomics methods due to their physicochemical properties and unfavorable amino acid sequences (e.g., membrane proteins, proteins inaccessible to commonly used proteases, etc.). However, due to the use of complementary proteases and extensive fractionation, we achieved exceptionally high individual protein coverage in this study (45% on average), which is higher than results from similar proteogenomics approaches (28–31). Therefore, we presume that a large fraction of the nonidentified proteins is highly specific to individual tissues or only expressed under context-specific conditions, neither of which were addressed with our experimental setup.

This observation is further corroborated by the phylostratigraphic analysis of all *S. macrospora* proteins: while the identified fraction mainly populates the oldest PS (PS1 and -2), nonidentified proteins are overrepresented in the youngest PS. Since essential proteins usually are highly conserved and tend to be more abundant (e.g., structural proteins or ribosomal proteins), they are also easier to detect under common conditions. In contrast, evolutionarily young proteins tend to be more specialized and their expression restricted to specific conditions, rendering them more difficult to detect experimentally.

Although the idea of integrating LC-MS-based proteomics data into genome annotation seems obvious and was done several times even before the term proteogenomics was first suggested by Jaffe et al. (2), corresponding studies face eminent challenges. Foremost, accurate control of the true FDR of novel identifications is crucial and has to be correctly accounted for, as described by Nesvizhskii (1). We accounted for this in a 2-fold manner. First, only spectra not matching the 1% FDR threshold in the first search against the v3 protein database were taken for a second search against the 6-frame translation database of the *S. macrospora* genome. Thus, the vast majority of matches to known peptide sequences do not erroneously lower the threshold for putatively novel hits, which potentially would lead to an underestimation of the error rate. Second, the FDR of novel hits was calculated solely for this subset, avoiding distortions from the known peptides.

We demonstrate that the novel peptides not only match in scoring, mass deviation, and peptide length but also perfectly correlate with predicted retention times. This is not a parameter of the Percolator peptide evaluation workflow (32) and therefore represents an additional line of evidence. In addition, since FDR estimation upon *de novo* peptide sequencing is not trivial (33, 34), all potentially novel peptides identified by *de novo* peptide sequencing were added to a combined database alongside the validated refined protein sequences.

After spectrum matching, the results were grouped and FDRs were individually calculated for canonical and novel matches, accounting for potentially different characteristics of both classes. Similar approaches were recently applied successfully for the *de novo*-aided identification of proteasomal spliced peptides (35, 36). Since we started from stringently filtered *de novo* peptide sequences meeting an FDR of 5% (33), after class-specific FDR calculation only 5 out of 118 initially reported SAAVs did not satisfy the criteria, leading to a theoretical *de novo* FDR of 4.2%. Even the number of peptide spectrum matches (PSMs) supporting those SAAVs increased from 192 (*de novo*) to 416 (database search), since the latter approach is more tolerant of missing fragment ions.

Seventeen of those SAAVs could be further validated by PRM analysis, where the respective peptides showed the expected retention times and fragmentation behavior.

To the best of our knowledge, this study is the first to employ a *de novo* peptide sequencing-aided approach in the context of proteogenomics and RNA editing. Since in our case the flow of information originates from the proteomics side, we emphasize that this is the most literal implementation of an approach termed proteogenomics, while in the previous cases of genome- and transcriptome-aided peptide identification it would be more appropriate to speak of genoproteomics.

Arguably, the benefit of similar genoproteomics or proteogenomics reannotation endeavors is greatest for newly sequenced organisms (3, 30, 31). However, here we demonstrate that even well-annotated and well-studied organisms such as *S. macrospora* benefit from orthogonal proteomics input, especially concerning orphan genes and genes poorly conserved across taxa, since these tend to be missed by homology-based annotation approaches. Among the 104 newly identified genes, 23 were marked as incomplete, since they were found at scaffold borders and, thus, were missing their N- or C-terminal sequence information. Additional work is required to further consolidate the *S. macrospora* genome assembly.

New genes could have emerged through numerous processes at different evolutionary stages (37). In this context, phylostratigraphy is a widely used method that allows genome-wide investigation of gene formation despite certain inherent limitations and a potential bias concerning rapidly evolving genes (38). Here, we were able to phylostratigraphically classify newly identified proteins and found an accumulation of proteins of younger phylostrata. Peculiarly, the novel protein SMAC_12920.3, harboring a conserved glucosamine-6-P-*N*-acetyltransferase (GNAT) family *N*-acetyltransferase domain, was traced to PS11 (*Sordariomycetes*), while phylostratigraphy also revealed homologues in PS1 among a wide range of *Actinobacteria*. Considering that no match was found in interjacent PS, SMAC_12920.3 might be the product of horizontal gene transfer (HGT) during the course of adaptive radiation in the *Sordariomycetes* clade.

As described recently, the nematode *Hopolamina* acquired GNATs late in evolution horizontally from plant-pathogenic actinomycetes (39). The phylogenetic tree of the newly described FAM7 family of *N*-acetyltansferases indicates multiple possible events of HGT from bacteria to numerous eukaryotes and archaea. Interestingly, alignment and phylogenetic tree construction of the SMAC_12920.3 sequence with representatives of the three described classes of fungal GCN5-related *N*-acetyltransferases (histone acetyltransferase family 7 [FAM7] GNAT) show no clear membership to either of these classes; rather, it clusters more closely to actinobacterial GNATs (see Fig. S6 in the supplemental material). Similarly, the gene SMAC_02065.3, encoding a mannosyl-3-phosphoglycerate synthase, shows an analogous PS pattern and has been speculated to be a product of HGT from members of the thermophile *Actinomycetales* (40).

We identified peptides of proteins ascribed to 45 AS events. For 21 of them, both splice variants were detected with distinct peptides. Alternative transcript splicing was long known to be prevalent in animals, e.g., the estimated AS rate for human multiexon genes is thought to be >90%, but it has also been shown to be present in fungal species, with rates of up to 18% in *Cryptococcus neoformans* (41). For *Neurospora crassa*, a close relative and member of the *Sordariaceae*, for 12.7% or 1,245 of the genes in its current annotation, two or more mRNA splice isoforms exist (42). However, although it has been extensively assumed that AS contributes significantly to functional protein diversity and proteome complexity, seemingly just a minor portion of alternatively spliced transcripts are actually translated into protein (43).

In addition to a described occurrence of peculiar amino acid sequences found at splice sites that were unfavorable for LC-MS-based analysis (44), we did not expect to find equivalent evidence for the translation of AS mRNA variants into protein compared to their transcript abundances. However, these lower-abundance variants might be of even greater interest, since they might represent functional context-specific alternatives. This assumption is supported by an accumulation of these AS events in related fungal species. The newly identified splice variant of *mek2* potentially represents this case. The genomic sequence coding for both splice variants is conserved in *Neurospora* and *Podospora* species of the *Sordariaceae*, while more distant fungi only code for a long variant lacking the intron that is alternatively spliced in the before-mentioned species. Sequence analysis of the extended MEK-2 N terminus has resemblance to a protein kinase-like domain, hinting at a signaling option masked or exposed through AS. Overall, our findings have the potential to augment our understanding of both fungal AS and AS in general.

In conclusion, our advanced proteogenomics approach, with the combination of proteogenomics and *de novo* peptide sequencing analysis, in conjunction with Blast2GO and phylostratigraphy, not only provides a comprehensive view of the proteome of *S. macrospora* and insights into the functional understanding of this multicellular organism but also immensely enhances its genome annotation quality. This approach offers a powerful basis to further develop applications for investigating fundamentals in eukaryotic cellular differentiation and organismic development.

## MATERIALS AND METHODS

**Strains and culture conditions.** The previously sequenced *S. macrospora* wild-type strain (S48977) was obtained from our laboratory collection and crossed several times to spore color mutant fus (45) for regeneration purposes. A vigorously growing, fully fertile, black-spored strain (R19027) was chosen as a new isolate and used in the experiments described here. The strain was grown on cornmeal medium (46).

**Reagents.** Guanidine hydrochloride (GuHCl), iodoacetamide (IAA), and ammonium bicarbonate (ABC; $NH_4HCO_3$) were acquired from Sigma-Aldrich, Steinheim, Germany. Calcium chloride ($CaCl_2$) was purchased from Merck (Darmstadt, Germany), and dithiothreitol (DTT) was bought from Roche Diagnostics (Mannheim, Germany). Sequencing-grade modified trypsin and endoproteinase GluC were obtained from Promega (Madison, WI, USA), and all chemicals for ultrapure high-performance liquid chromatography (HPLC) solvents, i.e, formic acid (FA), trifluoroacetic acid (TFA), and acetonitrile (ACN), were acquired from Biosolve (Valkenswaard, Netherlands).

***De novo* assembly and transcriptome-based annotation of the *S. macrospora* wtR genome.** *S. macrospora* genomic DNA was prepared as previously described by pulverizing mycelium in liquid nitrogen and phenolization (47). A 380-bp insert library (2× 151-bp paired-end sequencing) was sequenced on an Illumina HiSeq 2500 at GATC (Constance, Germany). Illumina raw data were trimmed with custom-made Perl scripts to remove reads with undetermined bases and for trimming of low-quality bases (phred score of $<$10) from the 3′ end as described previously (11). Trimmed reads were assembled with SPAdes 3.10 (48) using k-mer lengths (-k) of 21, 33, 55, 77, 99, and 127. For annotation, available *S. macrospora* transcriptome data from total mycelia grown under different growth conditions (11, 17, 18) were *de novo* assembled using Trinity 2.4.0 (49). The *de novo*-assembled transcripts together with predicted transcripts from the previous *S. macrospora* annotation (11), as well as predicted proteins from *S. macrospora* and *N. crassa* (11, 50), were used for gene model predictions using MAKER 2.31.18 (51). *S. macrospora* locus tags (with the format SMAC_xxxxx) were mapped onto the predicted gene models based on BLAST results (52) using custom-made Perl scripts. Approximately 1,000 gene models were manually corrected based on mapped RNA-Seq data using the genome browser Artemis (53).

**Protein extraction.** For protein extraction of surface cultures, the *S. macrospora* wild type was precultured in petri dishes with liquid BMM, CM (46, 54), or SWG (derived from synthetic crossing medium [55] and containing $KNO_3$ [1 g/liter], $KH_2PO_4$ [1 g/liter], $MgSO_47H_2O$ [0.5 g/liter], NaCl [0.1 g/liter], $CaCl_2$ [0.1 g/liter], trace elements [0.1 ml/liter], arginine [1 g/liter], glucose [20 g/liter], soluble starch [40 g/liter; Difco], and biotin [0.1 mg/liter]) for 2 days at 27°C. Three standardized inoculates of BMM precultures were transferred into petri dishes with 20 ml liquid BMM and cultivated for 3, 5, and 7 days at 27°C and 40 rpm. For CM and SWG surface cultures, four and five inoculates were transferred into petri dishes with 20 ml liquid CM and SWG, respectively, and cultivated for 3 days at 27°C and 40 rpm. For shaking cultures, the precultures were cultivated in petri dishes with liquid BMM for 3 days at 27°C. One inoculate was transferred into 100-ml flasks with 80 ml liquid BMM and incubated for 2 days at 27°C, 100 rpm, and constant light (56). For cell wall lysis and protein extraction, dried mycelium was ground in liquid nitrogen, suspended in extraction buffer (50 mM Tris-HCl, pH 7.4, 250 mM NaCl, 10% glycerol, 0.05% NP-40, 1 mM phenylmethylsulfonyl fluoride, 0.2% protease inhibitor cocktail IV [Calbiochem], 1.3 mM benzamidine, 1% phosphatase inhibitor cocktails II and III [Sigma-Aldrich]), and centrifuged for 30 min at 4°C and 15,000 rpm. The supernatant was prepared for mass spectrometry.

**Determination of protein concentration and carbamidomethylation.** Protein concentration in all sample lysates was estimated by performing a calorimetric bicinchoninic acid assay (BCA protein concentration assay kit; Pierce) according to the manufacturer's protocol. Carbamidomethylation was performed by first reducing cysteines by addition of DTT to a final concentration of 10 mM and incubation for 30 min at 56°C. Free thiols were then alkylated with 30 mM IAA for 30 min at room temperature in the dark, and excess IAA was quenched by further addition of 10 mM DTT.

**Sample clean-up and proteolysis.** Prior to digestion, samples were purified by ethanol precipitation. A 10-fold excess of ice-cold ethanol was added to an aliquot of each sample, corresponding to 100 $\mu$g of total protein content, and incubated for 1 h at –40°C. Protein precipitates were centrifuged for

30 min at 12,000 rpm, 4°C, and after removal of the supernatant, pellets were washed with 200 $\mu$l ice-cold acetone. After centrifugation for 20 min at the same settings as those described above, supernatant was discarded and samples were dried under a laminar-flow hood. For digestion with trypsin, proteins were first resuspended in 20 $\mu$l 6 M GuHCl, followed by dilution with 50 mM ABC (pH 7.8) to reach a final concentration of 0.2 M the former. $CaCl_2$ was added at a concentration of 2 mM, and finally trypsin was added at a ratio of 1:20 (protease:substrate [wt/wt]). Digestions were carried out at 37°C for 14 h and stopped by the addition of 10% (vol/vol) TFA to a final concentration of 1%. Desalting of the peptides and assessment of the quality of the digests were done as described before (57). Samples dedicated to GluC digestion were resolubilized in 8 M urea, 50 mM ABC (pH 7.8) and diluted with an additional 50 mM ABC buffer to reach a final concentration of 0.4 M urea. Endoproteinase GluC was added at a 1:30 protease/substrate ratio (wt/wt), and samples were incubated for 14 h at 25°C. Acidification, desalting, and quality control were carried out as described above.

**Mass spectrometric analysis.** Samples for global analysis were fractionated after digestion and subjected to LC-MS/MS analysis using an Ultimate 3000 RSLCnano HPLC coupled to a Q Exactive HF mass spectrometer, while samples dedicated to label-free quantification and PRM analysis were measured unfractionated on an Ultimate 3000 RSLCnano HPLC coupled to an Orbitrap Fusion Lumos mass spectrometer (all from Thermo Scientific). Detailed settings are described in Text S1 in the supplemental material.

**Interpretation of MS/MS spectra.** For detailed information on the applied data interpretation workflows and database search settings, see Text S1. In short, *de novo* peptide sequencing was conducted by combined usage of the PEAKS Studio 7.5 (58), Novor (59), and pNovo+ (60) algorithms as described previously (33). Proteogenomic searches were conducted by consecutively searching the spectra against the *S. macrospora* genome (v3) protein database (PEAKS Studio 7.5 [58]) and the preprocessed 6-frame translation of the genome (v3) (Mascot 2.6.1), without and with allowed protein N-terminal acetylation. After every search step, matching spectra were excluded from the subsequent search. For appropriate FDR calculation of RNA editing variants, a hybrid database was generated, database searches were conducted with MS-GF+ (v10282) via the SearchGUI interface (version 3.2.20) (61), and FDR was calculated using the MSnID R package (62). The final comprehensive database search was conducted via Proteome Discoverer 2.2 (Thermo Fischer Scientific), and label-free quantification experiments were analyzed using Progenesis (version 3.0.6039; Nonlinear Dynamics, Newcastle upon Tyne, U.K.) in conjunction with SearchGUI (61), followed by statistical evaluation using R (v3.3.1) (63) base methods.

**Proteogenomic analysis.** Result files from searches against the *S. macrospora* genome were loaded into R (version 3.3.1) (63) using the mzID package. After compensating for redundancies introduced by the search approach by allowing for overlapping genome fragments, only unique peptides were kept for downstream analysis. To display the results in common genome browsers, identifications were converted to proBAM files as described by the HUPO Proteomics Standards Initiative (v 1.0.0). Identifications were first converted to SAM files and further sorted, indexed, and converted to BAM files using the free software tool SAMtools (64). Results from initial database searches against the predicted *S. macrospora* protein database were converted to proBAM similarly using the proBAMsuite R package (65).

Finally, peptide sequences resulting from the combined *de novo* peptide sequencing approach were first mapped to the known *S. macrospora* protein sequences using PepExplorer 2.0 (version 0.1.0.78) (66) with the minimum identity threshold set to 0.5, allowing peptides of 6 or more amino acids and not using the decoy tag option. All Ile amino acids in the database were first converted to Leu in order to compensate for the inability to discriminate between these two amino acids by this approach. Additionally, peptide sequences were aligned to the *S. macrospora* genome using the BLAST (version 2.3.4) (52) tBLASTn algorithm with the recommended parameters for short input sequences. Results of both alignments were filtered for unique hits with complete identity and converted to .bed format in order to be readable by common genome browsers. Annotation refinement was done in the Artemis genome browser (Wellcome Sanger Institute, release 16.0.0) (53). RNA-Seq data published previously (12) (GEO accession numbers GSM832531 to GSM832534; reads from wild-type sexual and vegetative mycelium and protoperithecia) were loaded alongside the proBAM files, and annotation refinement was performed by adhering to following rules. (i) If intergenic peptide hits were in disagreement with the existing annotation, (A) at least two distinct novel peptides had to be identified in order for the refinement to be carried out or (B) one novel peptide alongside at least one known peptide sequence was identified in case of refinements of splice sites or TIS. (ii) For intragenic peptide hits, novel gene annotations were reported as class I (high confidence) if (A) at least two distinct novel peptide sequences were detected and RNA-Seq depth was sufficiently high or as class II (medium confidence) if (B) one distinct peptide sequence was detected and RNA-Seq depth was sufficiently high or (C) two or more distinct novel peptides were identified and RNA-Seq depth was insufficient but homologous sequences were found in related organisms.

**SAAV identification.** Potential RNA editing sites were identified using a subset of the mapped *de novo* sequencing data, showing one or more amino acid mismatches to the translated genomic data. Prefiltering of potential SAAV peptides was done by discarding single-mismatch hits showing a theoretical mass difference of 42.01056 Da at the N-terminal position (Ser→Glu exchange not distinguishable from N-terminal acetylation) or a mass difference of 0.98402 Da at any position (Gln→Glu or Asn→Asp not distinguishable from deamidation). Nucleotide sequences of potential editing sites next were extracted, and the only peptides that were kept were those where A-to-I editing could theoretically occur. For identification of stop-loss editing events, peptides partially matching to known protein C termini and additionally comprising a Trp residue were extracted from the data set. Further, all spectra were manually quality controlled and filtered, excluding identifications which could be explained equally well by posttranslational modifications of the relevant or neighboring amino acids.

**Validation of potential SAAVs and stop-loss events with PRM.** PRM runs were analyzed using Skyline (version 4.1.0; MacCoss Lab Software, USA) (67). The top 5 transitions for each canonical and modified peptide were selected, and diagnostic transitions specific for the SAAVs in question were monitored where appropriate. This was done to further discriminate putative true amino acid exchanges from isobaric posttranslational modifications on neighboring amino acids. Similarity of the monitored transitions to the library spectra was ensured by using a dot product cutoff of 0.85.

**Phylostratigraphy analysis.** Construction of a phylostratigraphic map of all *S. macrospora* protein sequences (10,004 after proteogenomics refinement) was performed as previously described (68, 69), using a perl pipeline (https://github.com/AlexGa/Phylostratigraphy) and the BLAST algorithm (v 2.2.31+) (52). A set of 15 phylostrata (PS) was defined based on the NCBI taxonomy browser, ranging from all cellular organisms (PS1) to *S. macrospora* (PS15). Analysis was performed against a target database provided by Drost et al. (68) (http://msbi.ipbhalle.de/download/phyloBlastDB_Drost_Gabel_Grosse_Quint.fa.tbz), with a BLASTp E value cutoff of 1e−5. Each gene was assigned to the oldest PS producing at least one hit below the E value cutoff. In cases where no hit was found, the gene was assigned to the youngest PS.

**Data availability.** The proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository under the data set identifier PXD014240. The *S. macrospora* wtR genome sequence (BioProject no. PRJNA391581) was deposited at DDBJ/ENA/GenBank under the accession number NMPR00000000. The version described here is version NMPR01000000. Illumina reads were deposited in the NCBI SRA database under accession number SRR5749461.

## SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at https://doi.org/10.1128/mBio.02367-19.

**TEXT S1**, DOCX file, 1.5 MB.

**FIG S1**, PDF file, 1.8 MB.

**FIG S2**, PDF file, 1.9 MB.

**FIG S3**, PDF file, 0.1 MB.

**FIG S4**, PDF file, 0.1 MB.

**FIG S5**, PDF file, 0.02 MB.

**FIG S6**, PDF file, 0.1 MB.

**TABLE S1**, PDF file, 0.04 MB.

**DATASET S1**, XLSX file, 0.1 MB.

**DATASET S2**, XLSX file, 0.03 MB.

## ACKNOWLEDGMENTS

We thank R. Krampe and S. Schlewinski for excellent technical support.

This study was funded by the Deutsche Forschungsgemeinschaft (Bonn, Germany) (KU517/11-2, NO407/7-1, SI835/6-1, SI835/8-2, and TE977/2-1). The funding agency was not involved in designing the experiments, conducting the work, generating and analyzing the data, or the decision to write this paper and submit it for publication.

B.B.L., I.T., R.M., and M.N. conceived and designed the experiments. R.M. performed the biological sample preparation. B.B.L. performed peptide-level separation and the LC-MS analysis. B.B.L. and A.S. designed the proteogenomics workflow. M.N. performed assembly and initial annotation. Analyses of data were performed by B.B.L., I.T., R.M., and M.N. B.B.L., I.T., M.N., and U.K. wrote the manuscript with assistance from R.M. and A.S. All authors were involved in the discussion of the manuscript and approved its final version.

## REFERENCES

1. Nesvizhskii AI. 2014. Proteogenomics: concepts, applications and computational strategies. Nat Methods 11:1114–1125. https://doi.org/10.1038/nmeth.3144.
2. Jaffe JD, Berg HC, Church GM. 2004. Proteogenomic mapping as a complementary method to perform genome annotation. Proteomics 4:59–77. https://doi.org/10.1002/pmic.200300511.
3. Armengaud J, Trapp J, Pible O, Geffard O, Chaumot A, Hartmann EM. 2014. Non-model organisms, a species endangered by proteogenomics. J Proteomics 105:5–18. https://doi.org/10.1016/j.jprot.2014.01.007.
4. Ren Z, Qi D, Pugh N, Li K, Wen B, Zhou R, Xu S, Liu S, Jones AR. 2019. Improvements to the rice genome annotation through large-scale analysis of RNA-Seq and proteomics data sets. Mol Cell Proteomics 18:86–98. https://doi.org/10.1074/mcp.RA118.000832.
5. Ravikumar V, Nalpas NC, Anselm V, Krug K, Lenuzzi M, Šestak MS, Domazet-Lošo T, Mijakovic I, Macek B. 2018. In-depth analysis of *Bacillus subtilis* proteome identifies new ORFs and traces the evolutionary history of modified proteins. Sci Rep 8:17246. https://doi.org/10.1038/s41598-018-35589-9.
6. Zhu Y, Orre LM, Johansson HJ, Huss M, Boekel J, Vesterlund M, Fernandez-Woodbridge A, Branca RMM, Lehtiö J. 2018. Discovery of

coding regions in the human genome by integrated proteogenomics analysis workflow. Nat Commun 9:903. https://doi.org/10.1038/s41467-018-03311-y.

7. Mitchell NM, Sherrard AL, Dasari S, Magee DM, Grys TE, Lake DF. 2018. Proteogenomic re-annotation of *Coccidioides posadasii* strain Silveira. Proteomics 18:1700173. https://doi.org/10.1002/pmic.201700173.
8. Moolhuijzen P, See PT, Hane JK, Shi G, Liu Z, Oliver RP, Moffat CS. 2018. Comparative genomics of the wheat fungal pathogen *Pyrenophora tritici-repentis* reveals chromosomal variations and genome plasticity. BMC Genomics 19:279–279. https://doi.org/10.1186/s12864-018-5059-1.
9. Datta KK, Patil AH, Patel K, Dey G, Madugundu AK, Renuse S, Kaviyil JE, Sekhar R, Arunima A, Daswani B, Kaur I, Mohanty J, Sinha R, Jaiswal S, Sivapriya S, Sonnathi Y, Chattoo BB, Gowda H, Ravikumar R, Prasad T. 2016. Proteogenomics of *Candida tropicalis*–an opportunistic pathogen with importance for global health. OMICS 20:239–247. https://doi.org/10.1089/omi.2015.0197.
10. Zhu Y, Engström PG, Tellgren-Roth C, Baudo CD, Kennell JC, Sun S, Billmyre RB, Schröder MS, Andersson A, Holm T, Sigurgeirsson B, Wu G, Sankaranarayanan SR, Siddharthan R, Sanyal K, Lundeberg J, Nystedt B, Boekhout T, Dawson TL, Heitman J, Scheynius A, Lehtiö J. 2017. Proteogenomics produces comprehensive and highly accurate protein-coding gene annotation in a complete genome assembly of *Malassezia sympodialis*. Nucleic Acids Res 45:2629–2643.
11. Nowrousian M, Stajich JE, Chu M, Engh I, Espagne E, Halliday K, Kamerewerd J, Kempken F, Knab B, Kuo H-C, Osiewacz HD, Pöggeler S, Read ND, Seiler S, Smith KM, Zickler D, Kück U, Freitag M. 2010. De novo assembly of a 40 Mb eukaryotic genome from short sequence reads*: sordaria macrospora*, a model organism for fungal morphogenesis. PLoS Genet 6:e1000891. https://doi.org/10.1371/journal.pgen.1000891.
12. Teichert I, Wolff G, Kück U, Nowrousian M. 2012. Combining laser microdissection and RNA-seq to chart the transcriptional landscape of fungal development. BMC Genomics 13:511. https://doi.org/10.1186/1471-2164-13-511.
13. Pöggeler S, Nowrousian M, Teichert I, Beier A, Kück U. 2018. Fruiting-body development in Ascomycetes, p 1–56. *In* Anke T, Schuffler A (ed), Physiology and genetics, 2nd ed, vol 2. Springer, Cham, Switzerland.
14. Teichert I, Nowrousian M, Poggeler S, Kück U. 2014. The filamentous fungus *Sordaria macrospora* as a genetic model to study fruiting body development. Adv Genet 87:199–244. https://doi.org/10.1016/B978-0-12-800149-3.00004-4.
15. Kück U, Beier AM, Teichert I. 2016. The composition and function of the striatin-interacting phosphatases and kinases (STRIPAK) complex in fungi. Fungal Genet Biol 90:31–38. https://doi.org/10.1016/j.fgb.2015.10.001.
16. Kück U, Radchenko D, Teichert I. 1 May 2019. STRIPAK, a highly conserved signaling complex, controls multiple eukaryotic cellular and developmental processes and is linked with human diseases. Biol Chem https://doi.org/10.1515/hsz-2019-0173.
17. Dirschnabel DE, Nowrousian M, Cano-Dominguez N, Aguirre J, Teichert I, Kück U. 2014. New insights into the roles of NADPH oxidases in sexual development and ascospore germination in *Sordaria macrospora*. Genetics 196:729–744. https://doi.org/10.1534/genetics.113.159368.
18. Schumacher DI, Lütkenhaus R, Altegoer F, Teichert I, Kück U, Nowrousian M. 2018. The transcription factor PRO44 and the histone chaperone ASF1 regulate distinct aspects of multicellular development in the filamentous fungus *Sordaria macrospora*. BMC Genet 19:112. https://doi.org/10.1186/s12863-018-0702-z.
19. Krokhin OV. 2006. Sequence-specific retention calculator. Algorithm for peptide retention prediction in ion-pair RP-HPLC: application to 300- and 100-Å pore size C18 sorbents. Anal Chem 78:7785–7795. https://doi.org/10.1021/ac060777w.
20. Marchler-Bauer A, Bryant SH. 2004. CD-Search: protein domain annotations on the fly. Nucleic Acids Res 32:W327–W331. https://doi.org/10.1093/nar/gkh454.
21. Banuett F. 1998. Signalling in the yeasts: an informational cascade with links to the filamentous fungi. Microbiol Mol Biol Rev 62:249–274.
22. Liu H, Li Y, Chen D, Qi Z, Wang Q, Wang J, Jiang C, Xu J-R. 2017. A-to-I RNA editing is developmentally regulated and generally adaptive for sexual reproduction in *Neurospora crassa*. Proc Natl Acad Sci U S A 114:E7756–E7765. https://doi.org/10.1073/pnas.1702591114.
23. Teichert I, Dahlmann TA, Kück U, Nowrousian M. 2017. RNA editing during sexual development occurs in distantly related filamentous Ascomycetes. Genome Biol Evol 9:855–868. https://doi.org/10.1093/gbe/evx052.
24. Liu H, Wang Q, He Y, Chen L, Hao C, Jiang C, Li Y, Dai Y, Kang Z, Xu J-R. 2016. Genome-wide A-to-I RNA editing in fungi independent of ADAR enzymes. Genome Res 26:499–509. https://doi.org/10.1101/gr.199877.115.
25. Teichert I. 2018. Adenosine to inosine mRNA editing in fungi and how it may relate to fungal pathogenesis. PLoS Pathog 14:e1007231. https://doi.org/10.1371/journal.ppat.1007231.
26. Liu J, Wang D, Su Y, Lang K, Duan R, Wu Y, Ma F, Huang S. 2019. FairBase: a comprehensive database of fungal A-to-I RNA editing. Database 2019: baz018. https://doi.org/10.1093/database/baz018.
27. Crosthwaite SK, Dunlap JC, Loros JJ. 1997. Neurospora wc-1 and wc-2: transcription, photoresponses, and the origins of circadian rhythmicity. Science 276:763–769. https://doi.org/10.1126/science.276.5313.763.
28. Castellana NE, Shen Z, He Y, Walley JW, Cassidy CJ, Briggs SP, Bafna V. 2014. An automated proteogenomic method uses mass spectrometry to reveal novel genes in *Zea mays*. Mol Cell Proteomics 13:157–167. https://doi.org/10.1074/mcp.M113.031260.
29. Marx H, Hahne H, Ulbrich SE, Schnieke A, Rottmann O, Frishman D, Kuster B. 2017. Annotation of the domestic pig genome by quantitative proteogenomics. J Proteome Res 16:2887–2898. https://doi.org/10.1021/acs.jproteome.7b00184.
30. Prasad TS, Mohanty AK, Kumar M, Sreenivasamurthy SK, Dey G, Nirujogi RS, Pinto SM, Madugundu AK, Patil AH, Advani J, Manda SS, Gupta MK, Dwivedi SB, Kelkar DS, Hall B, Jiang X, Peery A, Rajagopalan P, Yelamanchi SD, Solanki HS, Raja R, Sathe GJ, Chavan S, Verma R, Patel KM, Jain AP, Syed N, Datta KK, Khan AA, Dammalli M, Jayaram S, Radhakrishnan A, Mitchell CJ, Na CH, Kumar N, Sinnis P, Sharakhov IV, Wang C, Gowda H, Tu Z, Kumar A, Pandey A. 2017. Integrating transcriptomic and proteomic data for accurate assembly and annotation of genomes. Genome Res 27: 133–144. https://doi.org/10.1101/gr.201368.115.
31. McAfee A, Harpur BA, Michaud S, Beavis RC, Kent CF, Zayed A, Foster LJ. 2016. Toward an upgraded honey bee (*Apis mellifera* L.) genome annotation using proteogenomics. J Proteome Res 15:411–421. https://doi.org/10.1021/acs.jproteome.5b00589.
32. Kall L, Canterbury JD, Weston J, Noble WS, MacCoss MJ. 2007. Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat Methods 4:923–925. https://doi.org/10.1038/nmeth1113.
33. Blank-Landeshammer B, Kollipara L, Biß K, Pfenninger M, Malchow S, Shuvaev K, Zahedi RP, Sickmann A. 2017. Combining de novo peptide sequencing algorithms, a synergistic approach to boost both identifications and confidence in bottom-up proteomics. J Proteome Res 16: 3209–3218. https://doi.org/10.1021/acs.jproteome.7b00198.
34. Muth T, Renard BY. 2018. Evaluating de novo sequencing in proteomics: already an accurate alternative to database-driven peptide identification? Brief Bioinform 19:954–970. https://doi.org/10.1093/bib/bbx033.
35. Mylonas R, Beer I, Iseli C, Chong C, Pak H-S, Gfeller D, Coukos G, Xenarios I, Müller M, Bassani-Sternberg M. 2018. Estimating the contribution of proteasomal spliced peptides to the HLA-I ligandome. Mol Cell Proteomics 17:2347–2357. https://doi.org/10.1074/mcp.RA118.000877.
36. Faridi P, Li C, Ramarathinam SH, Vivian JP, Illing PT, Mifsud NA, Ayala R, Song J, Gearing LJ, Hertzog PJ, Ternette N, Rossjohn J, Croft NP, Purcell AW. 2018. A subset of HLA-I peptides are not genomically templated: evidence for cis- and trans-spliced peptide ligands. Sci Immunol 3:eaar3947. https://doi.org/10.1126/sciimmunol.aar3947.
37. Neme R, Tautz D. 2014. Evolution: dynamics of de novo gene emergence. Curr Biol 24:R238–R240. https://doi.org/10.1016/j.cub.2014.02.016.
38. Moyers BA, Zhang J. 2015. Phylostratigraphic bias creates spurious patterns of genome evolution. Mol Biol Evol 32:258–267. https://doi.org/10.1093/molbev/msu286.
39. Noon JB, Baum TJ. 2016. Horizontal gene transfer of acetyltransferases, invertases and chorismate mutases from different bacteria to diverse recipients. BMC Evol Biol 16:74. https://doi.org/10.1186/s12862-016-0651-y.
40. Borges N, Jorge CD, Goncalves LG, Goncalves S, Matias PM, Santos H. 2014. Mannosylglycerate: structural analysis of biosynthesis and evolutionary history. Extremophiles 18:835–852. https://doi.org/10.1007/s00792-014-0661-x.
41. Grützmann K, Szafranski K, Pohl M, Voigt K, Petzold A, Schuster S. 2014. Fungal alternative splicing is associated with multicellular complexity and virulence: a genome-wide multi-species study. DNA Res 21:27–39. https://doi.org/10.1093/dnares/dst038.
42. Kersey PJ, Allen JE, Allot A, Barba M, Boddu S, Bolt BJ, Carvalho-Silva D, Christensen M, Davis P, Grabmueller C, Kumar N, Liu Z, Maurel T, Moore

B, McDowall MD, Maheswari U, Naamati G, Newman V, Ong CK, Paulini M, Pedro H, Perry E, Russell M, Sparrow H, Tapanari E, Taylor K, Vullo A, Williams G, Zadissia A, Olson A, Stein J, Wei S, Tello-Ruiz M, Ware D, Luciani A, Potter S, Finn RD, Urban M, Hammond-Kosack KE, Bolser DM, De Silva N, Howe KL, Langridge N, Maslen G, Staines DM, Yates A. 2018. Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species. Nucleic Acids Res 46:D802–D808. https://doi.org/10.1093/nar/gkx1011.

43. Tress ML, Abascal F, Valencia A. 2017. Alternative splicing may not be the key to proteome complexity. Trends Biochem Sci 42:98–110. https://doi.org/10.1016/j.tibs.2016.08.008.
44. Wang X, Codreanu SG, Wen B, Li K, Chambers MC, Liebler DC, Zhang B. 2018. Detection of proteome diversity resulted from alternative splicing is limited by trypsin cleavage specificity. Mol Cell Proteomics 17:422–430. https://doi.org/10.1074/mcp.RA117.000155.
45. Nowrousian M, Teichert I, Masloff S, Kück U. 2012. Whole-genome sequencing of *Sordaria macrospora* mutants identifies developmental genes. G3 2:261–270. https://doi.org/10.1534/g3.111.001479.
46. Esser K. 1982. Cryptogams: cyanobacteria, algae, fungi, lichens. Cambridge University Press, Cambridge, United Kingdom.
47. Traeger S, Altegoer F, Freitag M, Gabaldon T, Kempken F, Kumar A, Marcet-Houben M, Pöggeler S, Stajich JE, Nowrousian M. 2013. The genome and development-dependent transcriptomes of *Pyronema confluens*: a window into fungal evolution. PLoS Genet 9:e1003820. https://doi.org/10.1371/journal.pgen.1003820.
48. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477. https://doi.org/10.1089/cmb.2012.0021.
49. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29:644–652. https://doi.org/10.1038/nbt.1883.
50. Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, FitzHugh W, Ma L-J, Smirnov S, Purcell S, Rehman B, Elkins T, Engels R, Wang S, Nielsen CB, Butler J, Endrizzi M, Qui D, Ianakiev P, Bell-Pedersen D, Nelson MA, Werner-Washburne M, Selitrennikoff CP, Kinsey JA, Braun EL, Zelter A, Schulte U, Kothe GO, Jedd G, Mewes W, Staben C, Marcotte E, Greenberg D, Roy A, Foley K, Naylor J, Stange-Thomann N, Barrett R, Gnerre S, Kamal M, Kamvysselis M, Mauceli E, Bielke C, Rudd S, Frishman D, Krystofova S, Rasmussen C, Metzenberg RL, Perkins DD, Kroken S, Cogoni C, Macino G, Catcheside D, Li W, Pratt RJ, Osmani SA, DeSouza CPC, Glass L, Orbach MJ, Berglund JA, Voelker R, Yarden O, Plamann M, Seiler S, Dunlap J, Radford A, Aramayo R, Natvig DO, Alex LA, Mannhaupt G, Ebbole DJ, Freitag M, Paulsen I, Sachs MS, Lander ES, Nusbaum C, Birren B. 2003. The genome sequence of the filamentous fungus *Neurospora crassa*. Nature 422:859–868. https://doi.org/10.1038/nature01554.
51. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, Holt C, Sánchez Alvarado A, Yandell M. 2007. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res 18:188–196. https://doi.org/10.1101/gr.6743907.
52. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.
53. Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA. 2012. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. Bioinformatics 28:464–469. https://doi.org/10.1093/bioinformatics/btr703.
54. Nowrousian M, Masloff S, Pöggeler S, Kück U. 1999. Cell differentiation during sexual development of the fungus *Sordaria macrospora* requires ATP citrate lyase activity. Mol Cell Biol 19:450–460. https://doi.org/10.1128/MCB.19.1.450.
55. Davis RH, de Serres FJ. 1970. Genetic and microbiological research techniques for Neurospora crassa. Methods Enzymol 17:79–143. https://doi.org/10.1016/0076-6879(71)17168-6.
56. Steffens EK, Becker K, Krevet S, Teichert I, Kück U. 2016. Transcription factor PRO1 targets genes encoding conserved components of fungal developmental signaling pathways. Mol Microbiol 102:792–809. https://doi.org/10.1111/mmi.13491.
57. Burkhart JM, Premsler T, Sickmann A. 2011. Quality control of nano-LC-MS systems using stable isotope-coded peptides. Proteomics 11:1049–1057. https://doi.org/10.1002/pmic.201000604.
58. Zhang J, Xin L, Shan B, Chen W, Xie M, Yuen D, Zhang W, Zhang Z, Lajoie GA, Ma B. 2012. PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. Mol Cell Proteomics 11:M111.010587. https://doi.org/10.1074/mcp.M111.010587.
59. Ma B. 2015. Novor: real-time peptide de novo sequencing software. J Am Soc Mass Spectrom 26:1885–1894. https://doi.org/10.1007/s13361-015-1204-0.
60. Chi H, Chen H, He K, Wu L, Yang B, Sun R-X, Liu J, Zeng W-F, Song C-Q, He S-M, Dong M-Q. 2013. pNovo+: de novo peptide sequencing using complementary HCD and ETD tandem mass spectra. J Proteome Res 12:615–625. https://doi.org/10.1021/pr3006843.
61. Vaudel M, Barsnes H, Berven FS, Sickmann A, Martens L. 2011. SearchGUI: an open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. Proteomics 11:996–999. https://doi.org/10.1002/pmic.201000595.
62. Petyuk V, Gatto L. 2016. MSnID: utilities for exploration and assessment of confidence of LC-MSn proteomics identifications. R package version 1.18.1. https://bioconductor.org/packages/release/bioc/html/MSnID.html. Accessed 23 October 2017.
63. R Development Core Team. 2016. R: a language and environment for statistical computing, v3.3.1. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.
64. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The sequence alignment/Map format and SAMtools. Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352.
65. Wang X, Slebos RJC, Chambers MC, Tabb DL, Liebler DC, Zhang B. 2016. proBAMsuite, a bioinformatics framework for genome-based representation and analysis of proteomics data. Mol Cell Proteomics 15:1164–1175. https://doi.org/10.1074/mcp.M115.052860.
66. Leprevost FV, Valente RH, Lima DB, Perales J, Melani R, Yates JR, III, Barbosa VC, Junqueira M, Carvalho PC. 2014. PepExplorer: a similarity-driven tool for analyzing de novo sequencing results. Mol Cell Proteomics 13:2480–2489. https://doi.org/10.1074/mcp.M113.037002.
67. MacLean B, Tomazela DM, Shulman N, Chambers M, Finney GL, Frewen B, Kern R, Tabb DL, Liebler DC, MacCoss MJ. 2010. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. Bioinformatics 26:966–968. https://doi.org/10.1093/bioinformatics/btq054.
68. Drost H-G, Gabel A, Grosse I, Quint M. 2015. Evidence for active maintenance of phylotranscriptomic hourglass patterns in animal and plant embryogenesis. Mol Biol Evol 32:1221–1231. https://doi.org/10.1093/molbev/msv012.
69. Domazet-Loso T, Brajković J, Tautz D. 2007. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. Trends Genet 23:533–539. https://doi.org/10.1016/j.tig.2007.08.014.
70. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. 2012. Predicting the functional effect of amino acid substitutions and indels. PLoS One 7:e46688. https://doi.org/10.1371/journal.pone.0046688.