*Send Orders for Reprints to reprints@benthamscience.net*

*Current Genomics*, **2021**, *22*, 373-383


**RESEARCH ARTICLE**

# Improving the Genome Annotation of *Rhizoctonia solani* Using Proteogenomics

Jiantao Shu[1], Mingkun Yang[2], Cheng Zhang[1], Pingfang Yang[1], Feng Ge[2,*] and Ming Li[1,*]

[1]*State Key Laboratory of Biocatalysis and Enzyme Engineering, School of Life Sciences, Hubei University, Wuhan, 430062 China;* [2]*Key Laboratory of Algal Biology, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan 430072, China*

ARTICLE HISTORY

Received: January 29, 2021
Revised: April 26, 2021
Accepted: May 30, 2021

*DOI: 10.2174/1389202922666211011143957*

**Abstract:** ***Background*:** *Rhizoctonia solani* is a pathogenic fungus that causes serious diseases in many crops, including rice, wheat, and soybeans. In crop production, it is very important to understand the pathogenicity of this fungus, which is still elusive. It might be helpful to comprehensively understand its genomic information using different genome annotation strategies.

***Methods*:** Aiming toimprove the genome annotation of *R. solani*, we performed a proteogenomic study based on the existing data. Based on our study, a total of 1060 newly identified genes, 36 revised genes, 139 single amino acid variants (SAAVs), 8 alternative splicing genes, and diverse post-translational modifications (PTMs) events were identified in *R. solani* AG3. Further functional annotation on these 1060 newly identified genes was performed through homology analysis with its 5 closest relative fungi.

***Results*:** Based on this, 2 novel candidate pathogenic genes, which might be associated with pathogen-host interaction, were discovered. In addition, in order to increase the reliability and novelty of the newly identified genes in *R. solani* AG3, 1060 newly identified genes were compared with the newly published available *R. solani* genome sequences of AG1, AG2, AG4, AG5, AG6, and AG8. There are 490 homologous sequences. We combined the proteogenomic results with the genome alignment results and finally identified 570 novel genes in *R. solani*.

***Conclusion*:** These findings extended *R. solani* genome annotation and provided a wealth of resources for research on *R. solani*.

**Keywords:** *Rhizoctonia solani*, proteogenomics, genomic annotation, pathogenic genes, post-translational modifications, SAAVs.

## 1. INTRODUCTION

Throughout their life cycle, plants are often exposed to different environmental stresses due to their sessile nature [1]. One of the serious environmental stresses is pathogen infection, which causes tremendous biological constraints on plant growth and productivity. Among all the plant pathogens, *Rhizoctonia solani* (*R. solani*) is an extremely harmful pathogenic fungus in the form of hyphae and sclerotia that resides in crop residues and soil [2, 3]. *R. solani* is a globally distributed soil-borne fungal plant pathogen [4, 5]. It has been a focal research topic since the 1960s as it is one of the most serious fungal pathogens to many staple crops production [6, 7]. There are at least 13 hyphal anastomosis groups (AG1 to AG13) of *R. solani* [4], which are species complex of genetically distinct fungi groups [8]. Among them, *R. solani* AG3 could affect cereal and legume crops, which are quite susceptible to *R. solani* AG3 disease [9, 10]. Previous studies have provided evidence of specific genomic sequences. As an example, the genome of *R. solani* was sequenced in 2014, which is approximately 51,705,945 bp long and consists of 326 scaffolds [11], including 308 "effect-like" genes that may be involved in plant pathogenicity [12]. There are numerous systematic studies of *R. solani* [5, 13, 14], but the molecular mechanisms of *R. solani* infection and pathogenesis are quite complicated to be fully elucidated [6, 15]. Quite a few studies have confirmed the ability of *R. solani* to produce toxins that have pathogenic effects on the host. However, the composition of the toxin has not been identified [16]. A previous study has indicated that pathogenic genes play a crucial role in the pathogenesis of *R. solani* [16]. The discovery of novel pathogenicity-related genes will not only pinpoint the pathogenic mechanism of pathogens, but also make use of novel pathogenicity-related genes to provide an effective way for biological control and breeding control in crop production [6, 7].

*Address correspondence to these authors at the State Key Laboratory of Biocatalysis and Enzyme Engineering, School of Life Sciences, Hubei University, Wuhan, 430062 China; Key Laboratory of Algal Biology, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan 430072, China; E-mail: limit@hubu.edu.cn

**1389-2029/21 $65.00+.00** 

There are three strategies for genomic annotation. The first one employs sequence tag-based database search [17]. The second uses homologous alignment annotation. The third strategy is the use of *de novo* sequencing [18]. However, genomic annotation based on these strategies is not accurate and complete [19-21]. This is shown by the fact that many peptides contain mutations or alternative splice forms, and many are not even present in any reference database at all. These facts indicate the existence of novel protein coding genes [22]. Consequently, it is reasonable to combine the nucleic acid data and proteome data in discovering novel genes or new events in the existing genes, which will optimize the current version of gene annotation.

With the advancement in nucleotide sequencing and mass spectrometry technologies, unprecedented high--throughput data could be generated at both nucleotide and protein levels. Due to this, proteogenomics has gradually expanded into the integration of proteomic, transcriptomic, and genomic data to identify gene coding sequences and protein body conditions in various cases [23, 24] and is widely applied in the scientific community. There are large-scale human proteogenomic studies that have been performed to identify many novel peptides and peptide variants [25, 26] and contributed to the major developments and breakthroughs in cancer research [27, 28]. GAPE tool could be used to analyze any sequenced eukaryotic genome, providing new insights into the biology of sequenced organisms. For example, the GAPE tool lays the foundation and in--depth explanation for the analysis of cyanobacteria and aflatoxin genome function [29, 30]. Compared with the other proteogenomic methods, the GAPE tool is a simple and efficient proteogenomic software, which requires minimal operator intervention. Moreover, it has a more strict filtering strategy for new peptides and uses multiple search engines to improve the accuracy of identified new peptides [31, 32]. In addition, the database search space increases with the number of PTMs. GAPE tool integrates non-restrictive PTMs search strategy (MODa) to discover a large number of new PTMs. In this study, we conducted a proteogenomic analysis on *R. solan* through the use of the GAPE tool [29], aiming at the discovery of novel genes, alternative splice forms, SAAV, and PTMs. The results might be utilized to improve and refine the genome annotation of *R. solani* and provide a wealth of resources for further study on *R. solani* and its pathogenesis.

## 2. MATERIALS AND METHODS

### 2.1. Data Acquisition and Processing

Raw mass spectrometry data were extracted from the PRIDE repository with the dataset identifier PRIDE: PXD002806, which were converted to "mgf" format using the pXtract software (version 2.0). Protein sequence, genomic sequence, and GFF file of R. solani AG3 were downloaded from the NCBI website (https://www.ncbi.nlm.nih.gov-/genome, GenBank assembly accession: GCA_000695385.1) [33]. SRA data were downloaded from NCBI Sequence Read Archive (SRA) database (PRJNA371695) and converted into the FASTQ format from the SRA format using fastp (version 0.20.0) for quality control and data filtering of FASTQ files. Trinity (version 2.9.0) was used for RNA-Seq de novo assembly. The RNA-Seq reads were assembled as follows: Trinity --seqType fq --left reads_1.fq --right reads_2.fq --CPU 10 --max_memory 30G. Besides, the protein sequence, genomic sequence, and GFF file of R. solani AG1, AG3, AG4, AG5, AG6, and AG8 have all been released in 2021 [34-37] and downloaded from https://api.ncbi.nlm.nih.gov/datasets/v1alpha/genome/download.

### 2.2. Construction and Utilization of Proteogenomic Database for Peptide and Protein Identification

Three different search tools, X!Tandem [38], Comet [39], and MSGF$^{+}$ [40], were used to search and identify the protein database, six-frame database, three-frame database, and the MS/MS data, respectively. The peptides identified from three different search tools were integrated and then mapped to the protein database to obtain novel peptides and known peptides. Among all collected data, only those that could be searched by all the three tools and mapped into the same sequence were further processed. The novel peptides were mapped to the six-frame database and the three-frame database for the discovery of novel events, including new genes. Consequently, for the uneven distribution of false positives in the peptides that were identified by the proteogenomic analysis, target-decoy search strategy [41] was utilized, which has a rigorous filtration strategy (≤1% separation) to calculate the actual false discovery rate (FDR) of known and novel peptides. Furthermore, in order to avoid a broad search space in spectrum search, the information collected for each protein was stored in an index file containing protein data.

Afterward, compared with one or two search tools [19, 42], the GAPE tool was able to expand search scope and was more efficient and comprehensive in peptide searches by using three different search tools for analyzing all obtained data.

### 2.3. Bioinformatics Analysis

In this study, the distribution of all identified novel events was targeted at the genome of *R. solani* and visualized by DNAPlotter [43]. Functional annotation was based on the Blast2GO (version 5.2), which assigned GO terms to identified novel genes and a unified vocabulary of biological processes, molecular functions, and cellular components in biology through GO [44, 45]. DeepLoc-1.0 software [46] was employed to predict the subcellular localization of 1060 identified novel proteins. We used blast (version 2.11.0+) to compare 1060 identified new genes with the available *R. solani* genome and their gene sequences. We used the reciprocal blast of TBtools (version 0.66834) (https://github.-com/CJ-Chen/TBtools) to analyze the conservation of the identified novel protein of *R. solani* and extract gene sequences in the available *R. solani* genome. The location of all identified peptides and proteins in the genome of *R.*

*solani* were graphically depicted using Artemis and Integrative Genomics Viewer (IGV) software [47]. We performed preliminary data analysis through simple commands of Linux and further processing was done using custom python scripts and statistics program R. See the supplementary file for more information.

## 3. RESULTS

### 3.1. Identification of Previously Unidentified Events

To improve the genome annotation of *R. solani*, we conducted a proteogenomic analysis according to the GAPE workflow (Fig. **1**) as reported [29]. All the raw data were extracted from the publicly available database as described in M&M. The spectra raw data were searched against the constructed proteogenomic database with three search engines to identify the peptides. All the identified peptides were then mapped into known protein databases with BLASTP to distinguish known peptides and unique orphan peptides, with the unique orphan peptides having not been previously recognized as a protein before for *R. solani*. The unique orphan peptides were then mapped to the genome by BLAST. The peptides that could be mapped to unique locations in the genome were named genome search specific peptides (GSSPs). The GSSPs were then blasted against the genome again to discover new events, including identified novel gene, revision of annotated gene model (revised gene), single amino acid variant (SAAV), and alternative splicing (AS) gene. Based on the mapping results, a total of 5513 unique genes and 714 shared genes (defined as genes encoding the proteins determined only through the shared peptides) were identified (Table **1**). Among them, there were 4516 confirmed exons and 1060 newly identified genes from the known protein sequence of *R. solani* AG3 (Table **1**; Supplemental Table **1A**). In addition, 36 revised genes which combine several adjacent exons into one longer exon (Table **1**; Supplemental Table **1B**), 139 SAAVs (Table **1**; Supplemental Table **1C**, **1D** and **1E**) and 8 AS genes (Table **1**; Supplemental Table **1F** and **1G**) were identified. The 139 SAAVs included 134 SAAVs from annotated proteins, 4 novel SAAVs and 1 revised SAAV (Table **1**; Supplemental Table **1C**-**E**). The 8 AS genes included 7 novel AS genes and 1 revised AS gene (Table **1**; Supplemental Table **1F** and **G**). To show the distribution of these new events on the genome of *R. solani* AG3, the top 15 largest scaffolds were selected to visualize these new events (Fig. **2A**).

The main purpose of the current study is to identify the new events in the genome. When two or more unique GSSPs were mapped to the region (not annotated as genes) with overlapped sequence, they were defined as newly identified genes. As shown in Fig. **3A**, two unique orphan peptides were mapped to an intergenic region across nucleotides 92070-92348 on scaffold 11 (Fig. **3A**), showing the existence of a novel protein-encoding gene. In addition to the identification of novel genes, the data could also be used to correct the annotated genes as well. Here we showed an example of gene structure correction, of which two adjacent exons were combined into one longer exon (Fig. **3B**). In this corrected novel structure, we found a unique orphan peptide located in the intron region between exon 1 and exon 2, and another unique orphan peptide showing a little extension of exon 2 (Fig. **3B**). Based on the peptides data, a new splicing event was discovered (Fig. **3C**). In this event, a novel gene was identified in an intergenic region, in which there was no annotated splicing site. This new splicing site was also supported by the transcriptomic data. As for SAAVs, an example was shown here, in which a T to C mutation resulted in a valine to alanine variation (Fig. **3D**). To verify the reliability of these new events, we conducted a transcriptomic analysis with the RNA-seq data extracted from the database. The existence of most of these new events was also supported by the transcriptomic data (Fig. **3**).

**Table 1. A summary of the results of GAPE tool identification.**

| Type | | Number of genes |
|---|---|---|
| Unique genes | | 5513 |
| Shared genes | | 714 |
| Confirmed exons | | 4516 |
| Novel protein coding regions | | 1060 |
| Gene/protein extensions | | 36 |
| Alternative splicing proteins | Novel proteins | 7 |
| | Revised proteins | 1 |
| Single amino acid variant proteins | Novel mutated proteins | 4 |
| | Revised proteins | 1 |
| | Annotated proteins | 134 |

### 3.2. Structure and Function Analyses of Newly Identified Genes

To further understand the newly determined genes, we analyzed the structure and function features of these genes. Most of the novel proteins were less than 400 amino acids in length, with the average length being 145 aa (Fig. **4A**), which indicates that they were mainly encoded by short ORFs (Supplemental Table **2A**). Analysis of protein data shows that the length of the new protein is significantly different from that of the identified and unidentified proteins (Fig. **4A**; Supplemental Table **2B** and **2C**). GC contents of the newly identified genes are slightly lower than those of other known genes (Fig. **4B**; Supplemental Table **2D**). The frequency of start codon was calculated for the identified novel protein-encoding genes as well, which showed that ATG is the predominant one accounting for more than 60%. GTG and TTG were the second and third frequent ones (Fig. **4C**; Supplemental Table **2A**). Most of the identified novel proteins were identified with a sequence coverage between 20-50% (Fig. **4D**; Supplemental Table **2A**).

Subcellular localization analysis of the 1060 identified novel proteins shows that a significant portion of the proteins (534) are localized in the mitochondrion (Supplemental Table 2E). Other subcellular localizations were as follows: extracellular (187), nucleus (185), cytoplasm (95), plastid (40), Golgi apparatus (7), endoplasmic reticulum (5), cell membrane (4), peroxisome (3) (Supplemental Table **2E**). Gene Ontology analysis was conducted with Blast2GO software (version 5.2) to annotate these 1060 newly identified

genes as illustrated in Fig. (**5A**). We found that quite a number of identified novel proteins were fully annotated according to their biological process and molecular functions (Supplemental Table **3A-C**), which was indicative of the existence and importance of the identified novel proteins in *R. solani* indirectly.

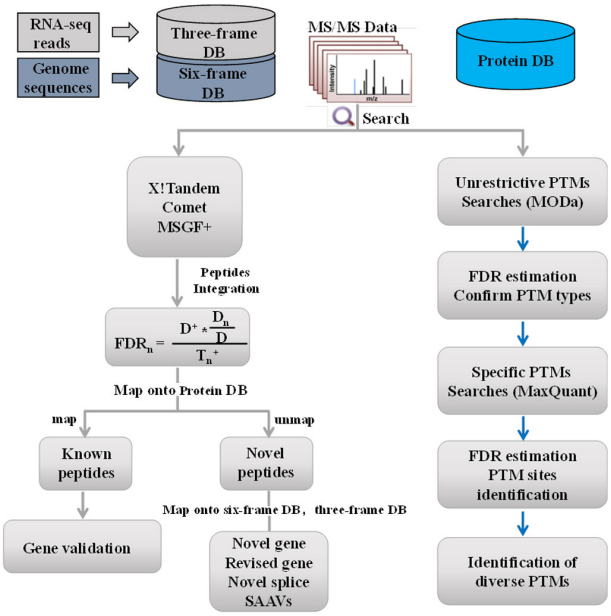


**Fig. (1). Overview of the workflow for the proteogenomic analysis of *R. solani*.** Protein DB: protein database fasta-formatted file containing all annotated protein sequences, Six-frame DB: six-frame translated genome database fasta-formatted file that created from the complete genomic sequence by this pipeline. Three-frame DB: three-frame translated genome database fasta-formatted file that created from the RNA-Seq reads by this pipeline. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

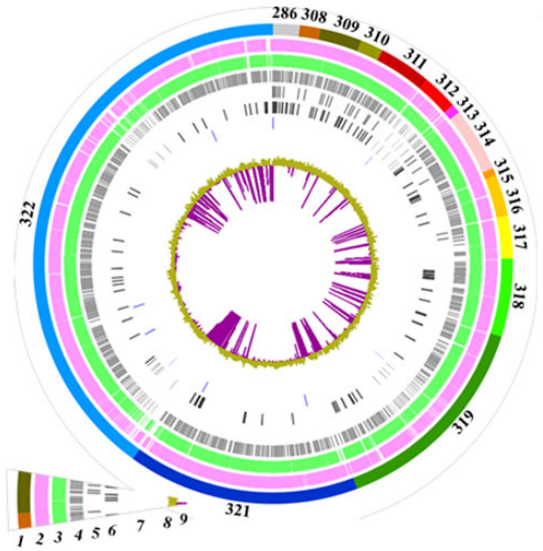


**Fig. (2). Proteome landscape of *R. solani***. Plots showing the distribution of identified novel events mapped to different scaffolds. The digital numbers out of the circle are the serial numbers of different scaffolds. Circles: 1, scaffolds; 2, genes on minus strand; 3, genes on plus strand; 4, newly identified genes; 5, revised genes; 6, SAAVs; 7, alternative splicing genes; 8, GC (>50%); 9, GC (<50%). This graph was drawn by DNAPlotter. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

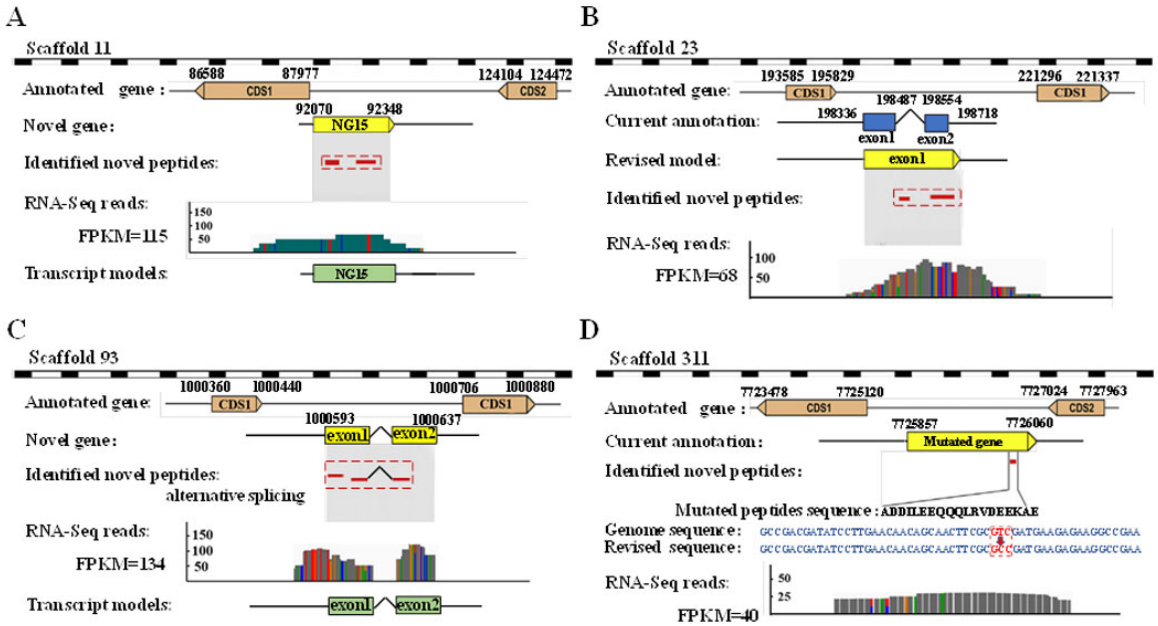


**Fig. (3). Identification of novel protein-coding genes, revised gene models, SAAVs and novel alternative splicing events.** (**A**) indicates two unique orphan peptides mapping to a region of *R. solani* genome scaffold 11, which was previously defined as the intergenic region. (**B**) indicates a revision on an annotated gene model, of which the previously annotated intron was partially mapped by two unique orphan peptides. (**C**) shows a novel alternative splicing event, of which three unique orphan peptides were mapped to intergenic regions with transcripts evidence supporting the existence of alternative splicing. (**D**) shows the identification of SAAVs on scaffold 311, in which an amino acid was mutated from to valine to alanine based on the unique orphan peptides information. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

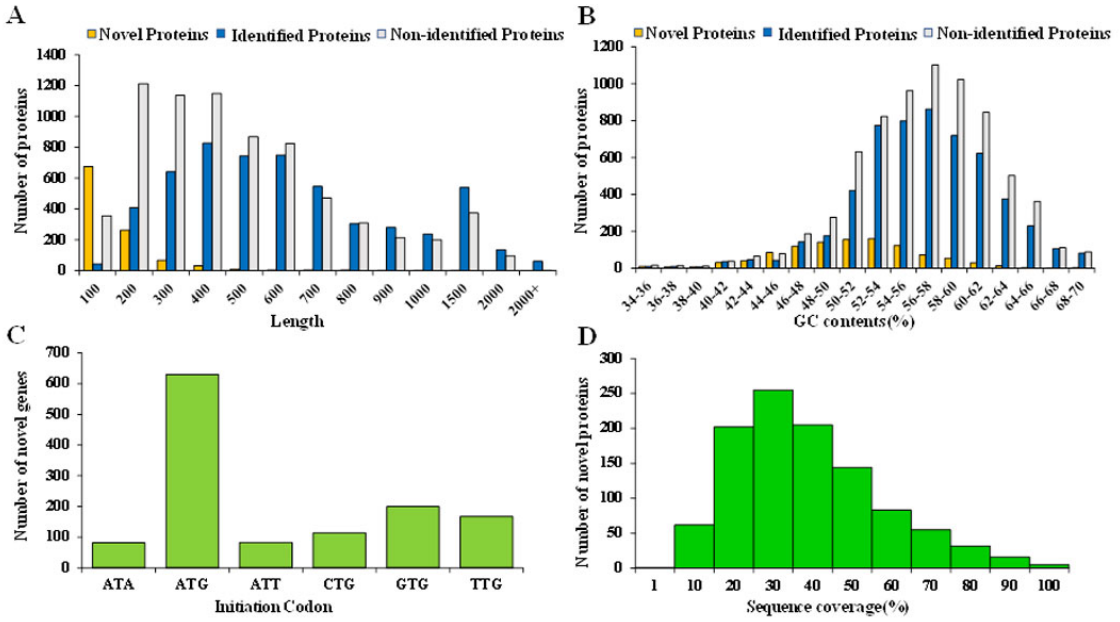


**Fig. (4). Summarization of proteomic data.** (**A**) Statistics of the length of all identified proteins in proteogenomic analyses. (**B**) Statistics of the GC content of all identified proteins in proteogenomic analyses. (**C**) Distribution of translation start codon of the coding genes of identified novel protein in proteogenomic analyses. (**D**) Statistics of protein sequence coverage. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

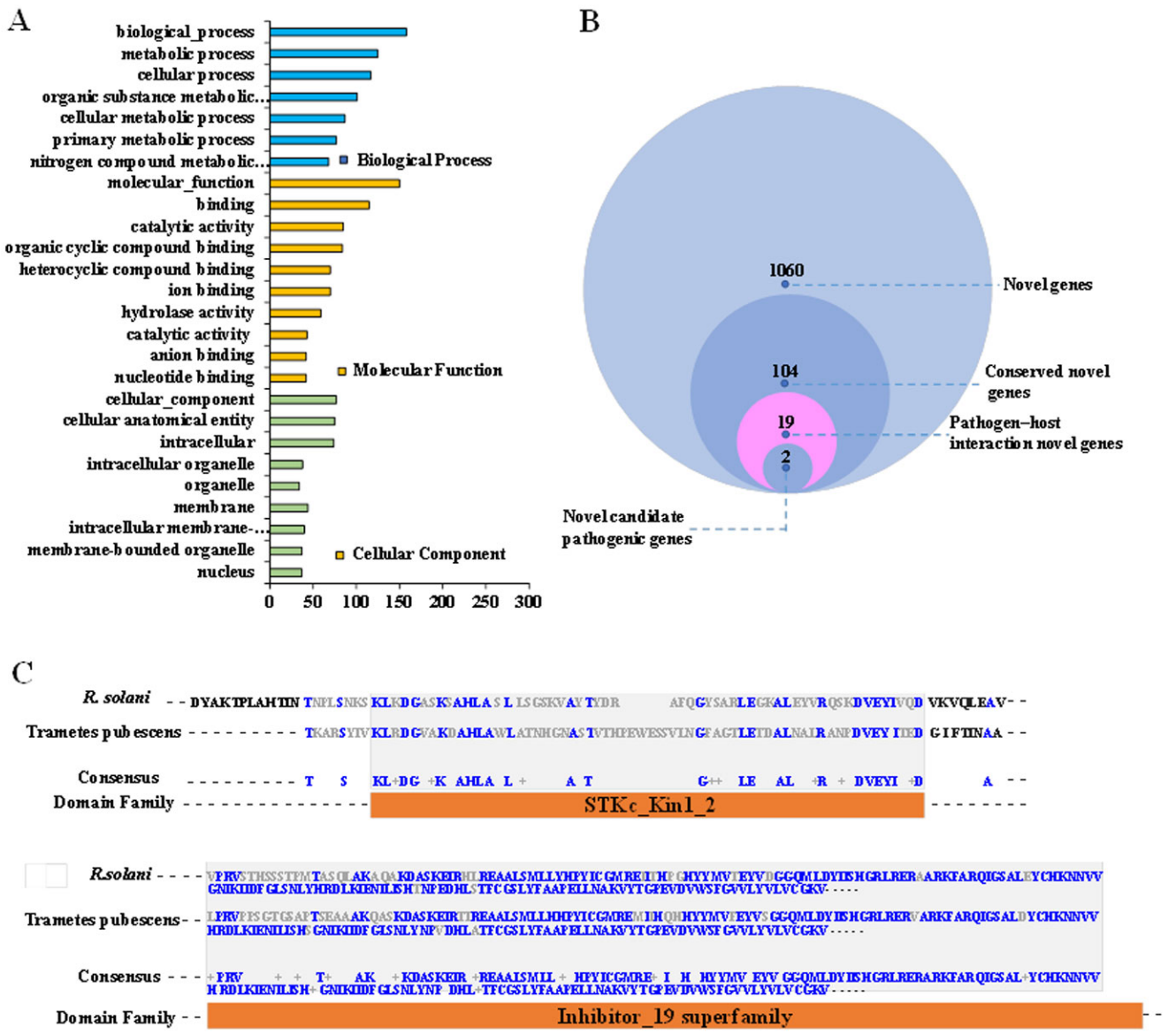


**Fig. (5). Functional analysis of novel protein-coding genes.** (**A**) Biological process, molecular function and cellular component annotation of the 1060 identified novel proteins by histogram. (**B**) Conservation analysis of 1060 identified novel *R. solani* proteins in other organisms. Through the homology analysis of the five closest species, we found two novel candidate pathogenic genes finally. (**C**) Sequence alignment of two novel candidate pathogenic proteins from *R. solani* with *Trametes pubescens*. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

In addition, the functional conservation of the identified novel proteins was also analyzed by comparing the identified novel proteins with their homologs from other five species showing closest phylogenetic relationship with *R. solani*. Among these identified novel protein-coding genes, there were 75, 63, 69, 73, and 72 newly identified genes, which are conserved in *Hypsizygus marmoreus*, *Leucoagaricus* sp. SymC.cos, *Phlebia centrifuga*, *Trametes pubescens*, and *Moniliophthora roreri*, respectively (Supplemental Table **3D-H**). After excluding the redundant genes, 104 conserved genes were detected, of which 19 are related to the interaction between host and pathogens (Fig. **5B**). These 19 identified new genes were analyzed on their conserved domains, which helped in discovering two novel candidate pathogenic genes in *R. solani* eventually (Fig. **5B**; Supplemental Table **3I**). As illustrated in Fig. **5C**, one novel candidate pathogenic protein has a homologous protein in *Trametes pubescens* and contains a STKc_Kin1_2 domain, which is the catalytic domain of Kin1, Kin2, and similar to Serine/Threonine Kinases. Kin1, whose orthologs are proper-

ly conserved among fungal pathogens, is essential for the pathogenesis and growth of phytopathogenic fungi [48, 49]. It suggests that this gene might be involved in the pathogenicity of *R. solani*. As for another candidate pathogenic protein, it has a homologous protein in *Trametes pubescens*, which contains an inhibitor_I9 superfamily related to cuticle-degrading protease and plays a key role as a virulence factor [50].

### 3.3. Comparison of Newly Identified Genes with Available *R. solani* Genomes

Based on the analysis results of the proteogenomic, we identified 1060 newly identified genes on the *R. solani* AG3 genome. In order to better explain the novelty of the newly identified genes, we compared these 1,060 identified novel gene sequences with the genomes sequence of different AG3 fusion groups. In the genome alignment results, AG1, AG2, AG4, AG5, AG6, and AG8 have 84, 278, 192, 206, 154, and 321 homologous sequences, respectively. After excluding the redundant genes, there are a total of 490 homologous sequences (Supplemental Table **3J**). In addition, we compared these 1,060 newly identified genes with the gene sequences of AG1, AG2, AG4, AG5, AG6, and AG8, and there are 380 homologous sequences. In the gene sequence alignment results, AG1, AG2, AG4, AG5, AG6, and AG8 have 65, 203, 157, 169, 133, and 194 homologous sequences, respectively. After excluding the redundant genes, there are a total of 380 homologous sequences (Supplemental Table **3K**). Obviously, the 490 homologous sequences in the genome alignment result are 110 genes more than the 380 homologous sequences in the gene sequence alignment. We regarded these 380 gene sequences alignments as known genes, and the remaining 110 genes in the intergenic region of the available *R. solani* genome were defined as unknown genes (Supplemental Table **3L**). As for the 570 gene sequences that do not match the available *R. solani* genome and gene sequences, we considered them to be novel genes (Supplemental Table **3M**).

The reciprocal blast results of 1060 identified new proteins with the protein sequences of AG1, AG2, AG4, AG5, AG6, and AG8 showed that there were 136, 129, 111,116, 114, and 134 protein sequences that were similar in sequence. After removing the repetitive sequences, there were 169 homologous sequences. If these 169 protein sequences are regarded as known protein sequences, 891 identified novel proteins remain (Supplemental Table **3N**). Comparing the homology of 891 identified novel proteins with the above five species, there were 12, 10, 9, 11, and 11 identified novel proteins, which are conserved in *Hypsizygus marmoreus*, *Leucoagaricus* sp. SymC.cos, *Phlebia centrifuga*, *Trametes pubescens*, and *Moniliophthora roreri*, respectively (Supplemental Table **3O**).

### 3.4. Discovery of Protein Post Translational Modifications in *R. solani*

Although many proteomic analyses have been conducted on *R. solani*, there is no protein PTMs study up-to-date. It is easier to understand the PTMs in eukaryotes by obtaining abundant PTMs results from mass spectrometry data cultivated under different conditions. In this study, we firstly applied MODa software to search the PTMs with the MS data. MODa has no limit on the number of modifications per peptide, which allows unlimited PTMs search. Based on the MODa searching, a total of 3038 unique peptides from 1111 proteins containing 4237 PTMs were detected (Supplemental Table **4A**). However, MODa could not pinpoint the modifications to the specific amino acid. Therefore, MaxQuant software (version 1.6.0.16) was applied for the second searching with the MS data to determine the modification occurring site. Based on the second searching, 16 typical PTMs were attached to the specific site (Fig. **6A**; Supplemental Table **4B**). To obtain reliable results, we compared the modified peptides and proteins lists from MODa and MaxQuant searching algorithms, which determined 695 modified peptides recognized by both software (Supplemental Table **4C**). To better understand the function of these modified proteins, COG analysis was conducted on these proteins (Supplemental Table **4D**). As shown in Fig. **6B**, the top five functional groups were coenzyme transport and metabolism, nucleotide transport and metabolism, replication, energy production and conversion, and cell wall/membrane/envelope biogenesis (Fig. **6B**).

## 4. DISCUSSION

With the advancement of sequencing technology, whole-genome sequencing has been an efficient method that could facilitate molecular genetic study. However, the genome assembly and annotation quality dramatically affect this efficiency. Specifically, many genes are annotated as predicted proteins because of lacking evidence from the protein level. Owing to the increasing sensitivity of mass spectrometer, many low abundant proteins could be detected, which provide more evidence of gene expression at the protein level. Using the MS data to refine the genome annotation has been developed into a strategy termed as “proteogenomics”. The proteogenomic strategy has been successfully applied in many different organisms [23, 24]. In this study, we applied our established proteogenomic pipeline [29] to refine the genome annotation in *R. solani*. A total of 1060 newly identified genes, 36 revised genes, 139 SAAVs, 8 alternative splicing genes, and diverse PTMs events were identified. However, the new genome sequences of *R. solani* from AG1 to AG13 have been published recently, and we compared our research results with them. There were 490 homologous sequences of *R. solani* from the 1,060 gene sequences identified as new genes. Eventually, we combined the proteogenomic results with the genome alignment results and identified 570 novel genes in *R. solani*.

Most of the identified novel proteins were short proteins, with the majority being less than 200 aa (Fig. **4**). As suggested by Yang *et al*. [29], they are defined as short proteins encoded by short ORFs, of which the typical size is about 100 aa. The important physiological function of short proteins was first proposed in the year of 2003 [51]. Since then, a lot of such short proteins with important functions havebeen

identified [52-54], based on which a database (http://www.sorfs.org) containing over thousands of short proteins was constructed [55]. In the current study, we identified more than 600 short proteins, among which 104 could be annotated through blasting against the database from the other five species. With the availability of genome information for more and more species, functions for these short proteins will be much better annotated.

Being a harmful pathogenic fungus for crops [2, 3], *R. solani* has been extensively studied since the 1960s [6, 7]. Among the 13 hyphal anastomosis groups (AG1 to AG13) of *R. solani* [4], *R. solani* AG3 could affect cereal and legume crops [9, 10]. Because of this, the genome of *R. solani* AG3 was sequenced and assembled into 326 scaffolds [11]. Although "effect-like" genes that may be involved in plant pathogenicity were identified [12], the molecular mechanisms of *R. solani* infection and pathogenesis still need to be elucidated [6, 15]. Obviously, the identification of pathogenic genes is helpful in understanding the pathogenesis of *R. solani* [16]. In this study, we identified 19 identified novel pathogen-host interaction genes containing 2 potential pathogenic genes. Further study on elucidating the function of these two pathogenic genes might be able to obtain a more in-depth understanding of the pathogenesis, which will undoubtedly facilitate the control of this damaged fungus. Proteogenomic analysis can be applied for the discovery of PTMs events. In our study, we combined MODa software (version 1.23) and MaxQuant software [56], searching to identify the existing PTMs and their sites precisely. How these PTMs contribute to the pathogenesis of *R. solani* is also very interesting and worth studying.

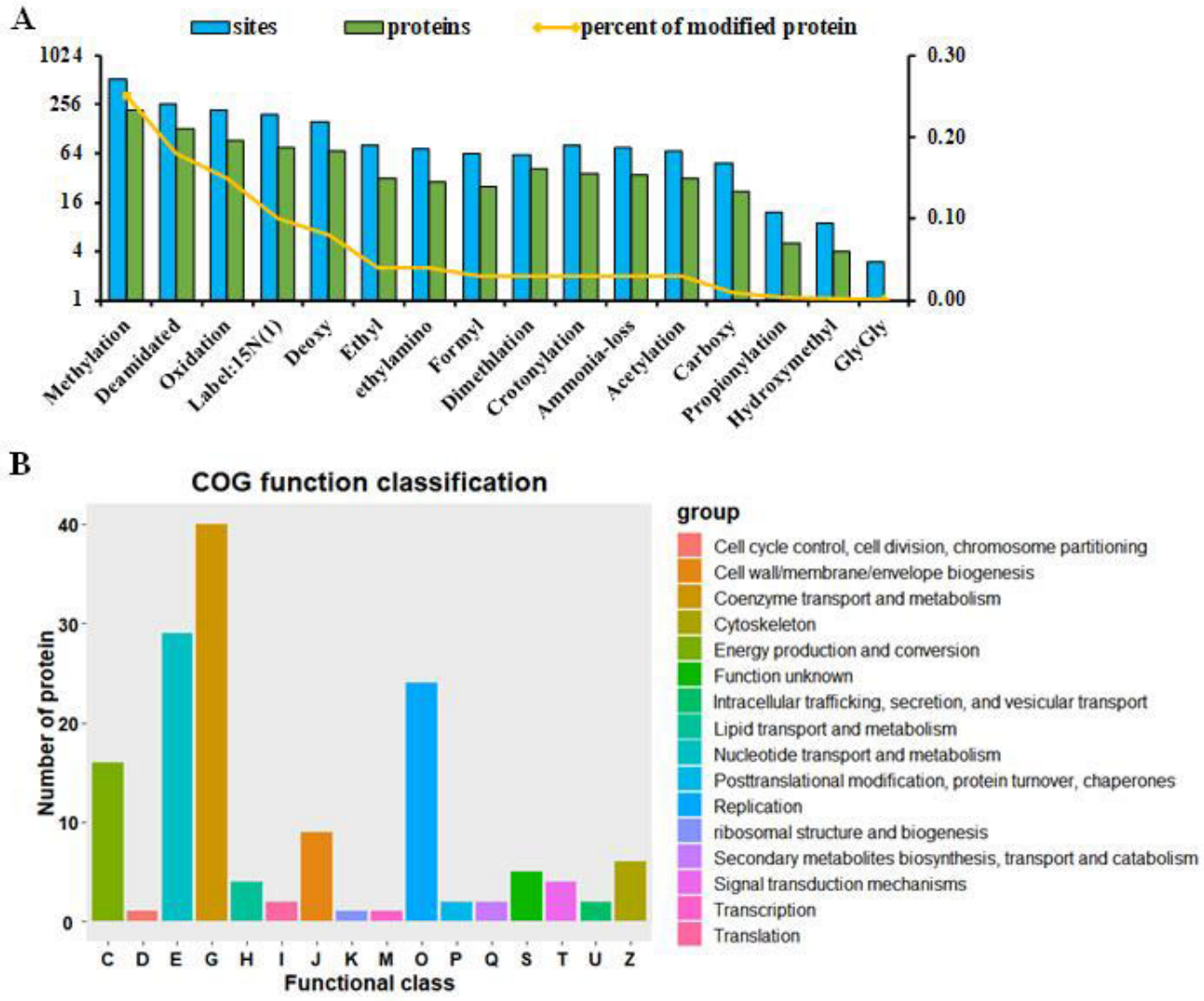


**Fig. (6). Overview of post-translational modifications of identified proteins.** (**A**) Statistics on the number of proteins and modification sites in different PTMs in *R. solani*. (**B**) Function categorization of the proteins with PTMs. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

## CONCLUSION

In summary, we corrected and refined previous gene annotations on the genome of *R. solani* AG3 through a proteogenomic strategy. Based on these analyses, two new candidate pathogenic genes were identified that are related to the pathogenic mechanism of *R. solani*. Some PTMs and their occurrence sites were also identified. Future studies elucidating the function of the new genes and PTMs events might be able to extend our understanding of the basic biology as well as the pathogenesis of *R. solani*.

## ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

## HUMAN AND ANIMAL RIGHTS

No animals/humans were used for studies that are the basis of this research.

## CONSENT FOR PUBLICATION

Not applicable.

## AVAILABILITY OF DATA AND MATERIALS

Data details available within article. see text "Protein sequence, genomic sequence, and GFF file of R. solani AG3 were downloaded from the NCBI website (https://www.ncbi.nlm.nih.gov/genome, GenBank assembly accession: GCA_000695385.1)".

## FUNDING

This work was supported by the National Natural Science Foundation of China (NSFC, No. 31271805).

## CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

## ACKNOWLEDGEMENTS

The authors are grateful to Dr. Rebecca Nejeri Damaris for proofreading and polishing the English language of the article.

## SUPPLEMENTARY MATERIAL

Supplementary material is available on the publisher's website along with the published material.

## REFERENCES

[1] Dodds, P.N.; Rathjen, J.P. Plant immunity: Towards an integrated view of plant-pathogen interactions. *Nat. Rev. Genet.,* **2010**, *11*(8), 539-548. http://dx.doi.org/10.1038/nrg2812 PMID: 20585331

[2] Kwon, Y.S.; Kim, S.G.; Chung, W.S.; Bae, H.; Jeong, S.W.; Shin, S.C.; Jeong, M.J.; Park, S.C.; Kwak, Y.S.; Bae, D.W.; Lee, Y.B. Proteomic analysis of *Rhizoctonia solani* AG-1 sclerotia maturation. *Fungal Biol.,* **2014**, *118*(5-6), 433-443. http://dx.doi.org/10.1016/j.funbio.2014.02.001 PMID: 24863472

[3] Dean, R.; Van Kan, J.A.; Pretorius, Z.A.; Hammond-Kosack, K.E.; Di Pietro, A.; Spanu, P.D.; Rudd, J.J.; Dickman, M.; Kahmann, R.; Ellis, J.; Foster, G.D. The Top 10 fungal pathogens in molecular plant pathology. *Mol. Plant Pathol.,* **2012**, *13*(4), 414-430. http://dx.doi.org/10.1111/j.1364-3703.2011.00783.x PMID: 22471698

[4] Lakshman, D.K.; Natarajan, S.S.; Lakshman, S.; Garrett, W.M.; Dhar, A.K. Optimized protein extraction methods for proteomic analysis of *Rhizoctonia solani*. *Mycologia,* **2008**, *100*(6), 867-875. http://dx.doi.org/10.3852/08-065 PMID: 19202841

[5] Shu, C.; Zhao, M.; Anderson, J.P.; Garg, G.; Singh, K.B.; Zheng, W.; Wang, C.; Yang, M.; Zhou, E. Transcriptome analysis reveals molecular mechanisms of sclerotial development in the rice sheath blight pathogen *Rhizoctonia solani* AG1-IA. *Funct. Integr. Genomics,* **2019**, *19*(5), 743-758. http://dx.doi.org/10.1007/s10142-019-00677-0 PMID: 31054140

[6] Wibberg, D.; Jelonek, L.; Rupp, O.; Hennig, M.; Eikmeyer, F.; Goesmann, A.; Hartmann, A.; Borriss, R.; Grosch, R.; Pühler, A.; Schlüter, A. Establishment and interpretation of the genome sequence of the phytopathogenic fungus *Rhizoctonia solani* AG1-IB isolate 7/3/14. *J. Biotechnol.,* **2013**, *167*(2), 142-155. http://dx.doi.org/10.1016/j.jbiotec.2012.12.010 PMID: 23280342

[7] Verwaaijen, B.; Wibberg, D.; Kröber, M.; Winkler, A.; Zrenner, R.; Bednarz, H.; Niehaus, K.; Grosch, R.; Pühler, A.; Schlüter, A. The *Rhizoctonia solani* AG1-IB (isolate 7/3/14) transcriptome during interaction with the host plant lettuce (Lactuca sativa L.). *PLoS One,* **2017**, *12*(5), e0177278. http://dx.doi.org/10.1371/journal.pone.0177278 PMID: 28486484

[8] Binder, M. The phylogenetic distribution of resupinate forms across the major clades of mushroom-forming fungi (Homobasidiomycetes). *Syst. Biodivers.,* **2005**, *3*, 113-157. http://dx.doi.org/10.1017/S1477200005001623

[9] Paulitz, T.C. Low input no-till cereal production in the pacific northwest of the U.S.: The challenges of root diseases. *Eur. J. Plant Pathol.,* **2006**, *115*, 271-281. http://dx.doi.org/10.1007/s10658-006-9023-6

[10] Anderson, J.P.; Singh, K.B. Interactions of arabidopsis and m. truncatula with the same pathogens differ in dependence on ethylene and ethylene response factors. *Plant Signal. Behav.,* **2011**, *6*(4), 551-552. http://dx.doi.org/10.4161/psb.6.4.14897 PMID: 21389781

[11] Cubeta, M.A.; Thomas, E.; Dean, R.A.; Jabaji, S.; Neate, S.M.; Tavantzis, S.; Toda, T.; Vilgalys, R.; Bharathan, N.; Fedorova-Abrams, N.; Pakala, S.B.; Pakala, S.M.; Zafar, N.; Joardar, V.; Losada, L.; Nierman, W.C. Draft genome sequence of the plant-pathogenic soil fungus *Rhizoctonia solani* anastomosis group 3 Strain rhs1ap. *Genome Announc.,* **2014**, *2*(5), 01072. http://dx.doi.org/10.1128/genomeA.01072-14 PMID: 25359908

[12] Hane, J.K.; Anderson, J.P.; Williams, A.H.; Sperschneider, J.; Singh, K.B. Genome sequencing and comparative genomics of the broad host-range pathogen *Rhizoctonia solani* AG8. *PLoS Genet.,* **2014**, *10*(5), e1004281. http://dx.doi.org/10.1371/journal.pgen.1004281 PMID: 24810276

[13] Okubara, P.A.; Dickman, M.B.; Blechl, A.E. Molecular and genetic aspects of controlling the soilborne necrotrophic pathogens rhizoctonia and pythium. *Plant Sci.,* **2014**, *228*, 61-70. http://dx.doi.org/10.1016/j.plantsci.2014.02.001 PMID: 25438786

[14] Zhang, J.; Chen, L.; Fu, C.; Wang, L.; Liu, H.; Cheng, Y.; Li, S.; Deng, Q.; Wang, S.; Zhu, J.; Liang, Y.; Li, P.; Zheng, A. Comparative transcriptome analyses of gene expression changes triggered by *rhizoctonia solani* AG1 IA infection in resistant and susceptible rice varieties. *Front. Plant Sci.,* **2017**, *8*, 1422. http://dx.doi.org/10.3389/fpls.2017.01422 PMID: 28861102

[15] Wibberg, D.; Jelonek, L.; Rupp, O.; Kröber, M.; Goesmann, A.; Grosch, R.; Pühler, A.; Schlüter, A. Transcriptome analysis of the phytopathogenic fungus *Rhizoctonia solani* AG1-IB 7/3/14 applying high-throughput sequencing of expressed sequence tags (ESTs). *Fungal Biol.,* **2014**, *118*(9-10), 800-813. http://dx.doi.org/10.1016/j.funbio.2014.06.007 PMID: 25209639

[16] Ghosh, S.; Kanwar, P.; Jha, G. Identification of candidate pathogenicity determinants of *Rhizoctonia solani* AG1-IA, which causes sheath blight disease in rice. *Curr. Genet.,* **2018**, *64*(3),

729-740.
http://dx.doi.org/10.1007/s00294-017-0791-7 PMID: 29196814

[17] Dasari, S.; Chambers, M.C.; Slebos, R.J.; Zimmerman, L.J.; Ham, A.J.; Tabb, D.L. TagRecon: High-throughput mutation identification through sequence tagging. *J. Proteome Res.,* **2010**, *9*(4), 1716-1726.
http://dx.doi.org/10.1021/pr900850m PMID: 20131910

[18] Ma, B.; Johnson, R. *De novo* sequencing and homology searching. *Mol Cell Proteomics.,* **2012**, *11*, 0111.014902.

[19] Prasad, T.S.; Harsha, H.C.; Keerthikumar, S.; Sekhar, N.R.; Selvan, L.D.; Kumar, P.; Pinto, S.M.; Muthusamy, B.; Subbannayya, Y.; Renuse, S.; Chaerkady, R.; Mathur, P.P.; Ravikumar, R.; Pandey, A. Proteogenomic analysis of candida glabrata using high resolution mass spectrometry. *J. Proteome Res.,* **2012**, *11*(1), 247-260.
http://dx.doi.org/10.1021/pr200827k PMID: 22129275

[20] Castellana, N.E.; Shen, Z.; He, Y.; Walley, J.W.; Cassidy, C.J.; Briggs, S.P.; Bafna, V. An automated proteogenomic method uses mass spectrometry to reveal novel genes in Zea mays. *Mol. Cell. Proteomics,* **2014**, *13*(1), 157-167.
http://dx.doi.org/10.1074/mcp.M113.031260 PMID: 24142994

[21] Agarwal, M.; Pathak, A.; Rathore, R.S.; Prakash, O.; Singh, R.; Jaswal, R.; Seaman, J.; Chauhan, A. Proteogenomic analysis of *burkholderia* species strains 25 and 46 isolated from uraniferous soils reveals multiple mechanisms to cope with uranium stress. *Cells,* **2018**, *7*(12), 269.
http://dx.doi.org/10.3390/cells7120269 PMID: 30545132

[22] Ruggles, K.V.; Krug, K.; Wang, X.; Clauser, K.R.; Wang, J.; Payne, S.H.; Fenyö, D.; Zhang, B.; Mani, D.R. Methods, tools and current perspectives in proteogenomics. *Mol. Cell. Proteomics,* **2017**, *16*(6), 959-981.
http://dx.doi.org/10.1074/mcp.MR117.000024 PMID: 28456751

[23] Helmy, M.; Sugiyama, N.; Tomita, M.; Ishihama, Y. Mass spectrum sequential subtraction speeds up searching large peptide ms/ms spectra datasets against large nucleotide databases for proteogenomics. *Genes Cells,* **2012**, *17*(8), 633-644.
http://dx.doi.org/10.1111/j.1365-2443.2012.01615.x PMID: 22686349

[24] Kumar, D.; Yadav, A.K.; Jia, X.; Mulvenna, J.; Dash, D. Integrated transcriptomic-proteomic analysis using a proteogenomic workflow refines rat genome annotation. *Mol. Cell. Proteomics,* **2016**, *15*(1), 329-339.
http://dx.doi.org/10.1074/mcp.M114.047126 PMID: 26560066

[25] Kim, M.S.; Pinto, S.M.; Getnet, D.; Nirujogi, R.S.; Manda, S.S.; Chaerkady, R.; Madugundu, A.K.; Kelkar, D.S.; Isserlin, R.; Jain, S.; Thomas, J.K.; Muthusamy, B.; Leal-Rojas, P.; Kumar, P.; Sahasrabuddhe, N.A.; Balakrishnan, L.; Advani, J.; George, B.; Renuse, S.; Selvan, L.D.; Patil, A.H.; Nanjappa, V.; Radhakrishnan, A.; Prasad, S.; Subbannayya, T.; Raju, R.; Kumar, M.; Sreenivasamurthy, S.K.; Marimuthu, A.; Sathe, G.J.; Chavan, S.; Datta, K.K.; Subbannayya, Y.; Sahu, A.; Yelamanchi, S.D.; Jayaram, S.; Rajagopalan, P.; Sharma, J.; Murthy, K.R.; Syed, N.; Goel, R.; Khan, A.A.; Ahmad, S.; Dey, G.; Mudgal, K.; Chatterjee, A.; Huang, T.C.; Zhong, J.; Wu, X.; Shaw, P.G.; Freed, D.; Zahari, M.S.; Mukherjee, K.K.; Shankar, S.; Mahadevan, A.; Lam, H.; Mitchell, C.J.; Shankar, S.K.; Satishchandra, P.; Schroeder, J.T.; Sirdeshmukh, R.; Maitra, A.; Leach, S.D.; Drake, C.G.; Halushka, M.K.; Prasad, T.S.; Hruban, R.H.; Kerr, C.L.; Bader, G.D.; Iacobuzio-Donahue, C.A.; Gowda, H.; Pandey, A. A draft map of the human proteome. *Nature,* **2014**, *509*(7502), 575-581.
http://dx.doi.org/10.1038/nature13302 PMID: 24870542

[26] Zhang, B.; Wang, J.; Wang, X.; Zhu, J.; Liu, Q.; Shi, Z.; Chambers, M.C.; Zimmerman, L.J.; Shaddox, K.F.; Kim, S.; Davies, S.R.; Wang, S.; Wang, P.; Kinsinger, C.R.; Rivers, R.C.; Rodriguez, H.; Townsend, R.R.; Ellis, M.J.; Carr, S.A.; Tabb, D.L.; Coffey, R.J.; Slebos, R.J.; Liebler, D.C. Proteogenomic characterization of human colon and rectal cancer. *Nature,* **2014**, *513*(7518), 382-387.
http://dx.doi.org/10.1038/nature13438 PMID: 25043054

[27] Zhang, Y. Al. a Pan-cancer proteogenomic atlas of pi3k/akt/mtor pathway alterations. *Cancer Cell.,* **2017**, *31*, 820-832.

[28] Gao, Q. Integrated proteogenomic characterization of hbv-related hepatocellular carcinoma. *Cell.,* **2019**, *179*, 561-577.

[29] Yang, M.; Lin, X.; Liu, X.; Zhang, J.; Ge, F. Genome annotation of a model diatom *Phaeodactylum tricornutum* using an integrated proteogenomic pipeline. *Mol. Plant,* **2018**, *11*(10), 1292-1307.
http://dx.doi.org/10.1016/j.molp.2018.08.005 PMID: 30176371

[30] Yang, M.; Zhu, Z.; Zhuang, Z.; Bai, Y.; Wang, S.; Ge, F. Proteogenomic characterization of the pathogenic fungus *Aspergillus flavus* reveals novel genes involved in aflatoxin production. *Mol. Cell. Proteomics,* **2020**, *20*, 100013.
http://dx.doi.org/10.1074/mcp.RA120.002144 PMID: 33568340

[31] Karpova, M.A.; Karpov, D.S.; Ivanov, M.V.; Pyatnitskiy, M.A.; Chernobrovkin, A.L.; Lobas, A.A.; Lisitsa, A.V.; Archakov, A.I.; Gorshkov, M.V.; Moshkovskii, S.A. Exome-driven characterization of the cancer cell lines at the proteome level: The NCI-60 case study. *J. Proteome Res.,* **2014**, *13*(12), 5551-5560.
http://dx.doi.org/10.1021/pr500531x PMID: 25333775

[32] Wen, B.; Xu, S.; Zhou, R.; Zhang, B.; Wang, X.; Liu, X.; Xu, X.; Liu, S. PGA: An R/Bioconductor package for identification of novel peptides using a customized database derived from RNA-Seq. *BMC Bioinformatics,* **2016**, *17*(1), 244.
http://dx.doi.org/10.1186/s12859-016-1133-3 PMID: 27316337

[33] Anderson, J.P.; Hane, J.K.; Stoll, T.; Pain, N.; Hastie, M.L.; Kaur, P.; Hoogland, C.; Gorman, J.J.; Singh, K.B. Mass-spectrometry data for *Rhizoctonia solani* proteins produced during infection of wheat and vegetative growth. *Data Brief,* **2016**, *8*, 267-271.
http://dx.doi.org/10.1016/j.dib.2016.05.042 PMID: 27331100

[34] Lee, D.Y.; Jeon, J.; Kim, K.T.; Cheong, K.; Song, H.; Choi, G.; Ko, J.; Opiyo, S.O.; Correll, J.C.; Zuo, S.; Madhav, S.; Wang, G.L.; Lee, Y.H. Comparative genome analyses of four rice-infecting *Rhizoctonia solani* isolates reveal extensive enrichment of homogalacturonan modification genes. *BMC Genomics,* **2021**, *22*(1), 242.
http://dx.doi.org/10.1186/s12864-021-07549-7 PMID: 33827423

[35] Wibberg, D.; Andersson, L.; Tzelepis, G.; Rupp, O.; Blom, J.; Jelonek, L.; Pühler, A.; Fogelqvist, J.; Varrelmann, M.; Schlüter, A.; Dixelius, C. Genome analysis of the sugar beet pathogen *Rhizoctonia solani* ag2-2iiib revealed high numbers in secreted proteins and cell wall degrading enzymes. *BMC Genomics,* **2016**, *17*, 245.
http://dx.doi.org/10.1186/s12864-016-2561-1 PMID: 26988094

[36] Zhang, Z.; Xia, X.; Du, Q.; Xia, L.; Ma, X.; Li, Q.; Liu, W. Genome sequence of *Rhizoctonia solani* anastomosis group 4 strain rhs4ca, a wide spread pathomycete in field crops. *Mol. Plant Microbe Interact.,* **2021**.
http://dx.doi.org/10.1094/MPMI-12-20-0362-A PMID: 33646817

[37] Mat Razali, N.; Hisham, S.N.; Kumar, I.S.; Shukla, R.N.; Lee, M.; Abu Bakar, M.F.; Nadarajah, K. Comparative genomics: Insights on the pathogenicity and lifestyle of *Rhizoctonia solani*. *Int. J. Mol. Sci.,* **2021**, *22*(4), 2183.
http://dx.doi.org/10.3390/ijms22042183 PMID: 33671736

[38] Brosch, M.; Swamy, S.; Hubbard, T.; Choudhary, J. Comparison of mascot and x!tandem performance for low and high accuracy mass spectrometry and the development of an adjusted mascot threshold. *Mol. Cell. Proteomics,* **2008**, *7*(5), 962-970.
http://dx.doi.org/10.1074/mcp.M700293-MCP200 PMID: 18216375

[39] May, D.H.; Tamura, K.; Noble, W.S. Param-Medic: A tool for improving ms/ms database search yield by optimizing parameter settings. *J. Proteome Res.,* **2017**, *16*(4), 1817-1824.
http://dx.doi.org/10.1021/acs.jproteome.7b00028 PMID: 28263070

[40] Kim, S.; Pevzner, P.A.M.S-G.F. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.,* **2014**, *5*, 5277.
http://dx.doi.org/10.1038/ncomms6277 PMID: 25358478

[41] Zhang, J.; Yang, M.K.; Zeng, H.; Ge, F. GAPP: A proteogenomic software for genome annotation and global profiling of post-translational modifications in prokaryotes. *Mol. Cell. Proteomics,* **2016**, *15*(11), 3529-3539.
http://dx.doi.org/10.1074/mcp.M116.060046 PMID: 27630248

[42] Castellana, N.E.; Pham, V.; Arnott, D.; Lill, J.R.; Bafna, V. Template proteogenomics: Sequencing whole proteins using an imperfect database. *Mol. Cell. Proteomics,* **2010**, *9*(6), 1260-1270.
http://dx.doi.org/10.1074/mcp.M900504-MCP200 PMID:

20164058

[43] Carver, T.; Thomson, N.; Bleasby, A.; Berriman, M.; Parkhill, J. DNAPlotter: Circular and linear interactive genome visualization. *Bioinformatics,* **2009**, *25*(1), 119-120. http://dx.doi.org/10.1093/bioinformatics/btn578 PMID: 18990721

[44] Conesa, A.; Götz, S.; García-Gómez, J.M.; Terol, J.; Talón, M.; Robles, M. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics,* **2005**, *21*(18), 3674-3676. http://dx.doi.org/10.1093/bioinformatics/bti610 PMID: 16081474

[45] Götz, S.; García-Gómez, J.M.; Terol, J.; Williams, T.D.; Nagaraj, S.H.; Nueda, M.J.; Robles, M.; Talón, M.; Dopazo, J.; Conesa, A. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.,* **2008**, *36*(10), 3420-3435. http://dx.doi.org/10.1093/nar/gkn176 PMID: 18445632

[46] Almagro Armenteros, J.J.; Sønderby, C.K.; Sønderby, S.K.; Nielsen, H.; Winther, O. DeepLoc: Prediction of protein subcellular localization using deep learning. *Bioinformatics,* **2017**, *33*(21), 3387-3395. http://dx.doi.org/10.1093/bioinformatics/btx431 PMID: 29036616

[47] Thorvaldsdóttir, H.; Robinson, J.T.; Mesirov, J.P. Integrative genomics viewer (IGV): High-performance genomics data visualization and exploration. *Brief. Bioinform.,* **2013**, *14*(2), 178-192. http://dx.doi.org/10.1093/bib/bbs017 PMID: 22517427

[48] Mylonakis, E.; Idnurm, A.; Moreno, R.; El Khoury, J.; Rottman, J.B.; Ausubel, F.M.; Heitman, J.; Calderwood, S.B. *Cryptococcus neoformans* Kin1 protein kinase homologue, identified through a *Caenorhabditis elegans* screen, promotes virulence in mammals. *Mol. Microbiol.,* **2004**, *54*(2), 407-419. http://dx.doi.org/10.1111/j.1365-2958.2004.04310.x PMID: 15469513

[49] Luo, Y.; Zhang, H.; Qi, L.; Zhang, S.; Zhou, X.; Zhang, Y.; Xu, J.R. FgKin1 kinase localizes to the septal pore and plays a role in hyphal growth, ascospore germination, pathogenesis, and localization of tub1 beta-tubulins in fusarium graminearum. *New Phytol.,* **2014**, *204*(4), 943-954. http://dx.doi.org/10.1111/nph.12953 PMID: 25078365

[50] Cito, A.; Barzanti, G.P.; Strangi, A.; Francardi, V.; Zanfini, A.; Dreassi, E. Cuticle-degrading proteases and toxins as virulence markers of *Beauveria bassiana* (Balsamo) Vuillemin. *J. Basic Microbiol.,* **2016**, *56*(9), 941-948. http://dx.doi.org/10.1002/jobm.201600022 PMID: 27198125

[51] Kessler, M.M.; Zeng, Q.; Hogan, S.; Cook, R.; Morales, A.J.; Cottarel, G. Systematic discovery of new genes in the *Saccharomyces cerevisiae* genome. *Genome Res.,* **2003**, *13*(2), 264-271. http://dx.doi.org/10.1101/gr.232903 PMID: 12566404

[52] Anderson, D.M.; Anderson, K.M.; Chang, C.L.; Makarewich, C.A.; Nelson, B.R.; McAnally, J.R.; Kasaragod, P.; Shelton, J.M.; Liou, J.; Bassel-Duby, R.; Olson, E.N. A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. *Cell,* **2015**, *160*(4), 595-606. http://dx.doi.org/10.1016/j.cell.2015.01.009 PMID: 25640239

[53] Hellens, R.P.; Brown, C.M.; Chisnall, M.A.W.; Waterhouse, P.M.; Macknight, R.C. The emerging world of small ORFs. *Trends Plant Sci.,* **2016**, *21*(4), 317-328. http://dx.doi.org/10.1016/j.tplants.2015.11.005 PMID: 26684391

[54] Hsu, P.Y.; Benfey, P.N. Small but mighty: Functional peptides encoded by small orfs in plants. *Proteomics,* **2018**, *18*(10), e1700038. http://dx.doi.org/10.1002/pmic.201700038 PMID: 28759167

[55] Olexiouk, V.; Van Criekinge, W.; Menschaert, G. An update on sORFs.org: A repository of small ORFs identified by ribosome profiling. *Nucleic Acids Res.,* **2018**, *46*(D1), D497-D502. http://dx.doi.org/10.1093/nar/gkx1130 PMID: 29140531

[56] Na, S.; Bandeira, N.; Paek, E. Fast multi-blind modification search through tandem mass spectrometry. *Mol Cell Proteomics.,* **2012**, *11*, 010199. http://dx.doi.org/10.1074/mcp.M111.010199