**BERNARDO DO VALE ARAÚJO MELO**

# GENOMIC STUDIES IN *Phakopsora pachyrhizi* AND IN ITS HYPERPARASITE *Simplicillium lanosoniveum*

Thesis submitted to the Plant Pathology Graduate Program of the Universidade Federal de Viçosa in partial fulfillment of the requirements for the degree of *Doctor Scientiae*.

Adviser: Sérgio Hermínio Brommonschenkel

**VIÇOSA - MINAS GERAIS**
**2023**

**Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa**

T
M528g
2023

Melo, Bernardo do Vale Araújo, 1994-
Genomic studies in *Phakopsora pachyrhizi* and in its hyperparasite *Simplicillium lanosoniveum* / Bernardo do Vale Araújo Melo. – Viçosa, MG, 2023.
1 tese eletrônica (140 f.): il. (algumas color.).

Texto em inglês.
Inclui apêndices.
Orientador: Sérgio Hermínio Brommonschenkel.
Tese (doutorado) - Universidade Federal de Viçosa, Departamento de Fitopatologia, 2023.
Inclui bibliografia.
DOI: https://doi.org/10.47328/ufvbbt.2023.385
Modo de acesso: World Wide Web.

1. Fungos fitopatogênicos - Controle - Aspectos genéticos. 2. Fungos fitopatogênicos - Controle biológico. 3. Virulência (Microbiologia). 4. Genômica. I. Brommonschenkel, Sérgio Hermínio, 1962-. II. Universidade Federal de Viçosa. Departamento de Fitopatologia. Programa de Pós-Graduação em Fitopatologia. III. Título.

CDD 22. ed. 579.59

Bibliotecário(a) responsável: Alice Regina Pinto Pires CRB-6/2523

**BERNARDO DO VALE ARAÚJO MELO**

# GENOMIC STUDIES IN *Phakopsora pachyrhizi* AND IN ITS HYPERPARASITE *Simplicillium lanosoniveum*

Thesis submitted to the Plant Pathology Graduate Program of the Universidade Federal de Viçosa in partial fulfillment of the requirements for the degree of *Doctor Scientiae*.

APPROVED: May 26, 2023.

Assent:

Bernardo do Vale Araújo Melo

Bernardo do Vale Araújo Melo
Author

Sérgio Hermínio Brommonschenkel
Adviser

## ACKNOWLEDGEMENTS

To my parents, family, and friends for emotional support.

To my advisor Sergio for the opportunity and confidence

To my lab colleagues for the good moments and for the help in our experiments.

To the Federal University of Viçosa, for the opportunity to complete the postgraduate course.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

To the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), for granting the scholarship.

## ABSTRACT

MELO, Bernardo do Vale Araújo, D.Sc., Universidade Federal de Viçosa, May, 2023. **Genomic studies in *Phakopsora pachyrhizi* and in its hyperparasite *Simplicillium lanosoniveum***. Adviser: Sérgio Hermínio Brommonschenkel.

*Phakopsora pachyrhizi* and *P. meibomiae* are the etiological agents of Asian soybean rust (ASR) and American soybean rust, respectively. Asian soybean rust is the most important fungal disease of soybean in Brazil and the use of genetic resistance is one of the methods of control recommended for the ASR management. Currently, seven loci containing dominant genes that confer resistance to *Phakopsora pachyrhizi* (*Rpp*) were identified, but the complementary avirulence genes of the pathogen have not yet been cloned and characterized. The *P. pachyrhizi* genome complexity, the absence of sexual reproduction, the obligate biotrophy, and the absence of transformation protocols limit the gene mapping and functional studies with this fungus. Recent advances in the DNA sequencing technologies and tools for genome assembly opened new possibilities to study genome structure, organization, and function for complex genomes such as the *P. pachyrhizi* one, avoiding difficulties posed by the fungus' biology. In addition, the assembly of complete genomes and comparative analysis allow the identification and characterization of complex loci and evolutionary processes related to pathogenicity, virulence, and hyperparasitism. Although the hyperparasitism of *P. pachyrhizi* by *Simplicillium lanosoniveum* has already been very well documented from a microscopic point of view, associated genes, and molecular mechanisms are still unknown. Therefore, in this study, genomic analyses were used to: a) Identification of *P. pachyrhizi* candidate avirulence genes corresponding to resistance genes *Rpp1b* (*Avr1*) and *Rpp5* (*Avr5*); b) Characterization of the genetic and genomic structure of the Mating-type system in *Phakopsora* sp., *P. meibomiae*, and *P. pachyrhizi*; and c) Assembly and annotation of *S. lanosoniveum* genome and identification of molecular mechanisms possibly associated with hyperparasitism of *P. pachyrhizi*. Two candidate genes for *Avr1* and four *Avr5* of *P. pachyrhizi* predicted to encode secreted proteins were identified. The mating-type system in the *Phakopsora* species that were analyzed is heterothallic, possibly tetrapolar, and one hormone receptor protein with an atypical structure was identified in *P. pachyrhizi*. The genome of the mycoparasite *S. lanosoniveum* was sequenced and chromosome-level assembly was obtained. The annotation of the *S. lanosoniveum* genome revealed enzymes and secondary metabolites unique to this species that may be related to its parasitism on *P. pachyrhizi*. The genetic and genomic resources developed create new perspectives for cloning and characterization of

avirulence genes of *P. pachyrhizi*, genes of *S. lanosoniveum* that encode specific enzymes and secondary metabolites, and also expand the understanding of the importance of sexual reproduction in the reproductive biology of *P. pachyrhizi*.

Keywords: Avirulence genes. Mating-type. Comparative genomics. Genetic control. Biological control. *Rpp1b*. *Rpp5*.

## RESUMO

MELO, Bernardo do Vale Araújo, D.Sc., Universidade Federal de Viçosa, maio de 2023. **Genomic studies in *Phakopsora pachyrhizi* and in its hyperparasite *Simplicillium lanosoniveum***. Orientador: Sérgio Hermínio Brommonschenkel.

*Phakopsora pachyrhizi e P. meibomiae* são os agentes causais da ferrugem asiática e americana em leguminosas, respectivamente. A ferrugem é a principal doença fúngica da soja no Brasil e a utilização de variedades resistentes é uma das medidas recomendadas para o seu manejo integrado. Sete locos que contém genes dominantes que conferem resistência a *P. pachyrhizi* já foram identificados, mas nenhum gene de avirulência correspondente do patógeno foi clonado e caracterizado até o momento. A complexidade do genoma de *P. pachyrhizi*, a ausência de reprodução sexuada funcional e de protocolos de transformação, associados ao parasitismo obrigatório, limitam os estudos de mapeamento genético e a identificação e análise funcional de genes neste patógeno. Os avanços recentes nas tecnologias de sequenciamento de DNA e de algoritmos de montagem de genomas abriram novas possibilidades de estudo da estrutura, organização e função de genomas complexos como o de *P. pachyrhizi,* permitindo contornar alguns dos entraves relacionados à biologia do fungo. Além disso, a obtenção de genomas completos e análises comparativas permitem identificar e caracterizar locos complexos e processos evolutivos relacionados a patogenicidade, virulência e hiperparasitismo. Embora o hiperparasitismo de *P. pachyrhizi* por *Simplicillium lanosoniveum* já tenha sido muito bem documentado do ponto de vista microscópico, ainda não se conhece os genes e mecanismos moleculares associados. Portanto, neste estudo foram utilizadas ferramentas de análise genômica para: a) Identificar candidatos a genes de avirulência de *P. pachyrhizi* correspondentes aos genes de resistência *Rpp1b* (*Avr1*) e *Rpp5* (*Avr5*); b) Caracterizar a estrutura genética e genômica do sistema *Mating-type* em *Phakopsora* sp., *P. meibomiae* e *P. pachyrhizi*; e c) Montar e anotar o genoma do micoparasita *S. lanosoniveum* e identificar mecanismos moleculares possivelmente associados ao hiperparasitismo em *P. pachyrhizi*. Foram identificados seis genes candidatos a *Avr1* e *Avr5* de *P. pachyrhizi* preditos como codificadores de proteínas secretadas. O sistema de *mating-type* nas espécies de *Phakopsora* analisadas é heterotálico, possivelmente tetrapolar, sendo identificada a sequência codificadora de uma proteína receptora de hormônios com estrutura atípica em *P. pachyrhizi*. O genoma do micoparasita *S. lanosoniveum* foi sequenciado e montado em nível cromossômico. A anotação do seu genoma revelou enzimas e metabólitos

secundários únicos desta espécie que podem estar relacionados ao seu parasitismo em *P. pachyrhizi*. Os resultados obtidos nessa tese abrem novas perspectivas para a clonagem e caracterização de genes de avirulência de *P. pachyrhizi*, genes que codificam enzimas e metabólitos secundários específicos de *S. lanosoniveum*, além de ampliar a compreensão da importância da reprodução sexuada na biologia reprodutiva de *P. pachyrhizi*.

Palavras-chave: Genes de avirulência. Mating-type. Genômica comparativa. Controle genético. Controle biológico. *Rpp1b*. *Rpp5*.

**SUMMARY**

## 1. GENERAL INTRODUCTION

Asian soybean rust and American soybean rust are soybean diseases caused by *Phakopsora pachyrhizi* (Syd. & P. Syd.) and *P. meibomiae* (Arthur), respectively. Both diseases are present in Brazil, but *P. pachyrhizi* is the prevalent species in soybean (*Glycine max* L. Merrill) fields, causing an important fungal disease of the crop (Godoy et al., 2016; Zambolim et al., 2022). Asian soybean rust management integrates several techniques, including the application of fungicides, sanitary vacuum, the use of early cultivars, sowing at the beginning of the recommended season, and genetic resistance (Godoy et al., 2016; Zambolim et al., 2022).

Resistance is a characteristic that depends on the interaction between the plant and the microorganism. Compatible interactions result in pathogen development and disease occurrence, while incompatible ones result in healthy plants. Plant pathogens secret effector molecules in their hosts to evade plant basal defense, which can result in plant susceptibility or be recognized by Nucleotide-binding domain leucine-rich repeat proteins (NLR) in plants, triggering defense responses and resulting in incompatible interaction, being denominated Avr proteins (Jones & Dangl, 2006; Dangl et al., 2013). Pathogens can also evade host recognition with novel effectors or natural changes, caused by genetic variation in their avirulence (*avr*) genes (Badet & Croll, 2020).

The gene-for-gene theory suggests that each gene of the host conferring resistance has a corresponding avirulence gene in the pathogen (Flor, 1971). Seven loci containing *Resistance to Phakopsora pachyrhizi* (*Rpp*) genes were identified in soybean genotypes (*Rpp 1-7*) (Childs et al, 2018; Garcia et al., 2008, Hyrten et al., 2007, 2009, Li et al., 2012, Silva et al., 2008), but no avirulence gene of *P. pachyrhizi* was identified yet. *P. pachyrhizi* genome complexity, the absence of sexual reproduction, the obligate biotrophy, and the absence of transformation protocols limit the gene mapping and functional studies with this fungus (Lorrain et al., 2019).

Population studies and the recent genome sequences of three *P. pachyrhizi* isolates from Brazil indicate that the population in South America reproduces asexually and indicates the absence of sexual reproduction (Jorge et al., 2015; Darben et al., 2020; Gupta et al., 2023). The mating-type system controls sexual reproduction in fungi, but it has never been reported in *Phakopora* genus, and considering that no crossing also has been reported since the introduction of this pathogen in Brazil, it is not known if *P. pachyrhizi* has a functional system with different mating types.

*Simplicillium lanosoniveum* is a mycoparasite fungus that infects *P. pachyrhizi*, entering its uredospores by its germinative pore and absorbing the cellular content (Ward et al., 2011; Gauthier et al., 2014). Although this fungus already has been reported as a potential agent for biological control of the Asian soybean rust (Ward et al., 2012) and the hyperparasitism of *P. pachyrhizi* also has been documented microscopically (Gauthier et al., 2014), no molecular mechanism associated with the mycoparasite towards *P. pachyrhizi* parasitism has been identified yet.

Recent advances in DNA sequencing technologies, including long-read sequencing platforms and tools for genome assembly opened new possibilities to study genome structure, organization, and function for complex genomes such as the *P. pachyrhizi* one, avoiding difficulties related to the fungus' biology (Pollard et al., 2018). Using these new advances, three *P. pachyrhizi* isolates were assembled and published recently (Gupta et al., 2023).

Comparative genomics has been used to identify avirulence genes in rust pathogens (Chen et al., 2017), to characterize the mating-type structure in rust pathogens (Cuomo et al., 2017), and to compare the genomic features of mycoparasite species (Karlsson et al., 2017). Here, three chapters will be presented, where comparative genomics was used to compare *P. pachyrhizi* isolates and *Phakopsora* species infecting soybean to identify candidate avirulence genes (chapter one) and to characterize the genetic and genomic structure of the mating-type system in *Phakopsora* species infecting Fabaceae (chapter two). One complete reference genome for *S. lanosoniveum* was assembled and it was compared with the genomes of other related fungi with different niches to identify possible molecular mechanisms associated with the mycoparasitism on *P. pachyrhizi* (chapter three).

**2. REFERENCES**

Badet, T., & Croll, D. (2020). The rise and fall of genes: origins and functions of plant pathogen pangenomes. *Current opinion in plant biology*, 56, 65-73.

Chen, J., Upadhyaya, N. M., Ortiz, D., Sperschneider, J., Li, F., Bouton, C., ... & Dodds, P. N. (2017). Loss of *AvrSr50* by somatic exchange in stem rust leads to virulence for *Sr50* resistance in wheat. *Science*, 358 (6370), 1607-1610.

Childs, S. P., King, Z. R., Walker, D. R., Harris, D. K., Pedley, K. F., Buck, J. W., ... & Li, Z. (2018). Discovery of a seventh *Rpp* soybean rust resistance locus in soybean accession PI 605823. *Theoretical and Applied Genetics*, 131, 27-41.

Cuomo, C. A., Bakkeren, G., Khalil, H. B., Panwar, V., Joly, D., Linning, R., ... & Fellers, J. P. (2017). Comparative analysis highlights variable genome content of wheat rusts and divergence of the mating loci. *G3: Genes, Genomes, Genetics*, 7 (2), 361-376.

Dangl, J. L., Horvath, D. M., & Staskawicz, B. J. (2013). Pivoting the plant immune system from dissection to deployment. *Science*, 341 (6147), 746-751.

Darben, L. M., Yokoyama, A., Castanho, F. M., Lopes-Caitar, V. S., da Cruz Gallo de Carvalho, M. C., Godoy, C. V., ... & Marcelino-Guimarães, F. C. (2020). Characterization of genetic diversity and pathogenicity of *Phakopsora pachyrhizi* mono-uredinial isolates collected in Brazil. *European Journal of Plant Pathology*, 156, 355-372.

Garcia, A., Calvo, É. S., de Souza Kiihl, R. A., Harada, A., Hiromoto, D. M., & Vieira, L. G. E. (2008). Molecular mapping of soybean rust (*Phakopsora pachyrhizi*) resistance genes: discovery of a novel locus and alleles. *Theoretical and Applied Genetics*, 117 (4), 545-553.

Gauthier, N.W., Maruthachalam, K., Subbarao, K.V., Brown, M., Xiao, Y., Robertson, C.L., Schneider, R.W. (2014). Mycoparasitism of *Phakopsora pachyrhizi*, the soybean rust pathogen, by *Simplicillium lanosoniveum*. *Biological Control* 76, 87–94. https://doi.org/10.1016/j.biocontrol.2014.05.008

Godoy, C. V., Seixas, C. D. S., Soares, R. M., Marcelino-Guimarães, F. C., Meyer, M. C., & Costamilan, L. M. (2016). Asian soybean rust in Brazil: past, present, and future. *Pesquisa Agropecuária Brasileira*, 51, 407-421.

Gupta, Y. K., Marcelino-Guimarães, F. C., Lorrain, C., Farmer, A. D., Haridas, S., Ferreira, E. G. C., ... & van Esse, H. P. (2023). Major proliferation of transposable elements shaped the genome of the soybean rust pathogen *Phakopsora pachyrhizi*. *Nature Communications*, 14, 1835.

Flor, Harold H. (1971). Current status of the gene-for-gene concept. *Annual review of phytopathology*, v. 9, n. 1, p. 275-296.

Hyten, D. L., Hartman, G. L., Nelson, R. L., Frederick, R. D., Concibido, V. C., Narvel, J. M., & Cregan, P. B. (2007). Map location of the *Rpp1* locus that confers resistance to soybean rust in soybean. *Crop Science*, 47 (2), 837-838.

Hyten, D. L., Smith, J. R., Frederick, R. D., Tucker, M. L., Song, Q., & Cregan, P. B. (2009). Bulked segregant analysis using the GoldenGate assay to locate the *Rpp3* locus that confers resistance to soybean rust in soybean. *Crop Science*, 49 (1), 265-271.

Jones, J. D., & Dangl, J. L. (2006). The plant immune system. *Nature*, 444 (7117), 323-329.

Jorge, V. R., Silva, M. R., Guillin, E. A., Freire, M. C. M., Schuster, I., Almeida, A. M. R., & Oliveira, L. O. (2015). The origin and genetic diversity of the causal agent of Asian soybean rust, *Phakopsora pachyrhizi*, in South America. *Plant Pathology*, 64 (3), 729-737.

Karlsson, M., Atanasova, L., Jensen, D. F., & Zeilinger, S. (2017). Necrotrophic mycoparasites and their genomes. *Microbiology Spectrum*, 5 (2), 5-2.

Li, S., Smith, J. R., Ray, J. D., & Frederick, R. D. (2012). Identification of a new soybean rust resistance gene in PI 567102B. *Theoretical and Applied Genetics*, 125, 133-142.

Lorrain, C., Gonçalves dos Santos, K. C., Germain, H., Hecker, A., & Duplessis, S. (2019). Advances in understanding obligate biotrophy in rust fungi. *New Phytologist*, 222 (3), 1190-1206.

Pollard, M. O., Gurdasani, D., Mentzer, A. J., Porter, T., & Sandhu, M. S. (2018). Long reads: their purpose and place. *Human molecular genetics*, 27 (R2), R234-R241.

Silva, D. C., Yamanaka, N., Brogin, R. L., Arias, C. A., Nepomuceno, A. L., Di Mauro, A. O., ... & Abdelnoor, R. V. (2008). Molecular mapping of two loci that confer resistance to Asian rust in soybean. *Theoretical and Applied Genetics*, 117, 57-63.

Ward, N.A., Robertson, C.L., Chanda, A.K., Schneider, R.W. (2012). Effects of *Simplicillium lanosoniveum* on *Phakopsora pachyrhizi*, the soybean rust pathogen, and its use as a biological control agent. *Phytopathology*, 102, 749–760. https://doi.org/10.1094/PHYTO-01-11-0031

Ward, N.A., Schneider, R.W., Aime, M.C. (2011). Colonization of soybean rust sori by *Simplicillium lanosoniveum*. *Fungal Ecology* 4, 303–308.

Zambolim, L., Reis, E. M., Guerra, W. D., Juliatti, F. C., & Menten, J. O. M. (2022). Integrated Management of Asian Soybean Rust. *European Journal of Applied Sciences*. Vol, 10 (2).

# CHAPTER 1

## IDENTIFICATION OF CANDIDATE AVIRULENCE GENES FROM *Phakopsora pachyrhizi* USING COMPARATIVE GENOMICS

**ABSTRACT**

Asian soybean rust, caused by *Phakopsora pachyrhizi*, is the most important fungus disease of soybean in Brazil and the use of genetic resistance plants is one of the methods of this disease management. Nowadays, seven loci containing dominant resistance genes conferring *Resistance to Phakopsora pachyrhizi* (*Rpp*) were identified, but no corresponding avirulence gene of the pathogen has been cloned and characterized. The *P. pachyrhizi* genome complexity, the absence of sexual reproduction, its obligate biotrophy, and the absence of transformation protocols limit the use of map-based gene cloning approaches and functional studies. Recent advances in the DNA sequencing technologies and tools for genome assembly opened new possibilities to study genome structure, organization, and function for complex genomes, allowing to bypass some technical difficulties that have hampered gene cloning and characterization in rust fungi. Comparative genomics of related species and genome-wide association studies between contrasting genotypes of the same pathogen species allows to identify candidate genes associated with parasitism and virulence. The main objective of this study was to identify the *P. pachyrhizi* candidate avirulence genes cognate to *Rpp1b* (*Avr1*) and *Rpp5* (*Avr5*) using this approach. A curated database of *P. pachyrhizi* containing 785 candidate effector proteins was created based on the predicted proteome of complete sequenced and annotated *P. pachyrhizi* isolates available in the Mycocosm database. Comparing 37 *P. pachyrhizi* isolates containing virulent and avirulent phenotypes, two candidate *Avr1* and four candidate *Avr5* genes were identified and characterized *in silico*. The candidate effector genes (*cAvr1-PHPA79* and *cAvr5-138_913Kb*), previously identified, were transiently expressed in *Nicotiana benthamiana* for subcellular localization studies and/or immunity suppression assays. The proteins encoded by the multiple paralog genes of the candidate *cAvr1-PHPA79* displayed co-localization with nuclei and possibly with the cell membrane, but none of the candidates suppressed plant immunity or was able to induce resistance responses in a heterologous system. Overall, the comparative genome analysis contributes to delimiting some candidate avirulence genes. However, it is necessary improvements in the pipeline of analysis to increase the chance of finding true candidate genes.

Keywords: Soybean rust. Genome-wide association study. *Rpp1b*. *Rpp5*.

## 1. INTRODUCTION

Asian soybean rust (ASR), caused by biotrophic basidiomycete *Phakopsora pachyrhizi*, is one of the most severe soybean (*Glycine max*) diseases around the world. The pathogen has a short life cycle and a wide host range, infecting over 95 plant species (Bonde et al., 2008), especially in the Fabaceae family, and high production of asexual spores that are easily dispersed to long distances by wind (Bromfield & Hartwig, 1980; Koch & Hoppe, 1983). These characteristics make ASR control more difficult, demanding a constant search for new approaches that could assist and facilitate disease management. In soybean plants, ASR causes intense defoliation and affects the grain's weight, reducing production and sometimes causing crop failure (Webb & Fellers, 2006; Goellner et al., 2010), due to losses that can exceed 80% of production without effective disease management.

Plants do not exhibit an adaptive defense system or even mobile cells, however, can respond to pathogens during infections (Jones & Dangl, 2006). Using the innate immune system plants can detect microorganisms based on the recognition of pathogen- or microbe-associated molecular patterns (PAMP/MAMP) by pattern recognition receptors (PRRs) localized in the plant cell surface (Rocafort et al., 2020). This first, broad-spectrum layer of defense is named PAMP-Triggered Immunity (PTI) (Schwessinger & Zipfel, 2008). To supplant the PTI, pathogens secret effector molecules into their host cells, where they can elicit Effector-Triggered Susceptibility (ETS). Some effectors can be recognized by Nucleotide-binding domain Leucine-rich repeat (NLR) proteins in plants (being therefore referred to as Avr proteins), activating the second defense layer denominated Effector-Triggered Immunity (ETI), a faster and broader defense response, usually associated with a Hypersensitive Response (HR) (Jones & Dangl, 2006; Dangl et al., 2013). Pathogens can also evade ETI and re-establish ETS by secreting novel effectors or by natural changes, caused by genetic variation in their avirulence genes, avoiding the recognition of the effector molecules by NLR proteins (Ngou et al., 2021, Pruitt et al., 2021, Yuan et al., 2021). In addition, plants can also restore ETI through the evolution of new NLR genes (Jones & Dangl, 2006; Badet & Croll, 2020).

Genetic analyses of the soybean-*P. pachyrhizi* pathosystem have led to the identification of seven loci, *Rpp1* to *Rpp7* (Resistance to *P. pachyrhizi* – Rpp) providing varying degrees of resistance in soybean (Childs et al, 2018; Garcia et al., 2008; Hyrten et al., 2007, 2009; Li et al., 2012; Silva et al., 2008). Most of the genes associated with each locus have not been cloned, only *Rpp1* and *Rpp4* have candidate genes (Mayer et al., 2009; Morales

et al., 2013; Peddley et al., 2019). In locus *Rpp1*, two genes have already been identified, *Rpp1* and *Rpp1b*. Only *Rpp1* is able to provide immunity (complete absence of macroscopic symptoms), whereas, for the other known genes, red-brown (RB) foliar lesions are observed in response to avirulent isolates of *P. pachyrhizi* (Peddley et al. 2019).

Although candidate genes for some *Rpp* resistance genes have been identified, no corresponding *P. pachyrhizi* effector (*Avr* gene) has been identified and characterized so far. Effectors are pathogen molecules, transported into plant cells (Koeck et al., 2011) and used to promote infection, proliferation, and survival in the host, reprograming host signaling defense and manipulating host cell structures and functions to induce an environment favorable to disease development (Figueroa et al., 2021). Effectors evolve rapidly (Persoons et al., 2017), exhibit unknown functional domains, and can be species-specific (de Guillen et al., 2019). Determining their action mode, expression periods, or how they trigger or overcome plant defense response in hosts is a difficult task, especially for rust pathogens (Lorrain et al., 2019). Despite the difficulty to genetically manipulate rust fungi, especially because they are not cultivable and have dikaryotic spores with notable heterozygosity degree (Cuomo et al., 2017), many advances have been made to identify and characterize rust effectors with a possible ability to supplant resistance.

Recently advances such as whole-genome, transcriptome, and proteome sequencing associated with improved bioinformatics tools have opened new options for identifying novel effector genes (Lorrain et al., 2019; Prasad et al., 2019), which are target molecules of NLR proteins, and hence Avr proteins. Some features, such as a smaller size (less than 300 amino acids), a higher cysteine content (greater than 3%), low molecular weight, presence of secretion signals, high expression, or taxonomical specificity, have been used as a criteria to identify candidate effectors (de Carvalho et al., 2017; Sperschneider et al., 2018). Effector mining in *P. pachyrhizi* based on transcriptome allowed Link et al. (2014) and Kunjet et al. (2016) to identify 156 and 35 *P. pachyrhizi* candidate effector genes, respectively. Cooper et al. (2016) using proteomics tools mined 319 *P. pachyrhizi* candidate secreted effectors proteins (CSEP). The complete genome sequencing of three isolates of *P. pachyrhizi* was another important advance in the effector research of this fungus because it enables genome-wide association studies (Gupta et al., 2023). Associating genome analysis with phenotypic traits, it is possible to comprehend how specific genomic loci and variation in their sequences affect the fungus characteristics including virulence and fungicide resistance (Ma & Michailides, 2005; Tam et al, 2019; Figueroa et al., 2020).

This approach was used to successfully characterize avirulence genes of wheat rust pathogen *Puccinia graminis* f. sp. *tritici*, such as *AvrSr27* (Upadhyaya et al., 2020), *AvrSr35* (Salcedo et al., 2017), and *AvrSr50* (Chen et al., 2017). So far, this approach has not been used in genomics studies of *P. pachyrhizi* aiming to isolate avirulence genes. Considering that the avirulence genes are generally dominant (Dodds et al., 2020) and it was possible to purify virulent *P. pachyrhizi* isolates from avirulent isolates (unpublished data), it was hypothesized that the avirulent isolates should be heterozygous at *Avr1* and *Avr5* loci.

Then, the objective of this study was to identify the candidate avirulence genes *Avr1* and *Avr5* of *P. pachyrhizi*, cognate to *Rpp1* and *Rpp5*, using comparative genomics. Two loci with *Avr1*candidate genes and four candidate genes for *Avr5* were identified and one of each candidate was submitted to functional studies in heterologous systems using *Nicotiana benthamiana*.

## 2. MATERIALS AND METHODS

### 2.1 Biological material

Monolesional *P. pachyrhizi* Syd. & P. Syd. isolate 71G2 was obtained from single lesions of naturally infected soybean (*Glycine max* L. Merrill) from the Mato Grosso state. One single lesion was detached and reinoculated on healthy detached leaves to purify the isolate and reproduce the disease symptoms. Uredospores were collected from the inoculated leaves and used to multiply the isolate on healthy soybean plants.

The soybean plants used in the spore multiplication were grown in 1L plastic pots in growth chambers at 22°C, with a photoperiod of 12/12 (light/dark). Thirty days old plants were inoculated with spore suspensions with $10^5$ uredospore $mL^{-1}$. The suspension was sprayed on leaves and the inoculated plants were kept in dew chambers at 25°C for 24h, then they were moved again to the growth chambers at 22°C. Uredospores were harvested with a spore collector 15 days after the inoculation. Collected uredospores were dehydrated in silica gel for 24h and preserved at -80°C.

The transient gene expression assays were performed with *Nicotiana benthamiana* obtained from the Sainsbury Laboratory (Norwich, UK). The plants were grown for 5-6 weeks in a growth chamber at 22°C with a 12/12 photoperiod. The infiltrated plants were kept in the same growth chamber for 48-72h to be analyzed.

### 2.2 DNA extraction and sequencing

The high molecular weight (HMW) genomic DNA from isolate 71G2 was extracted from spores using a modified CTAB protocol (Schewessinger & Rathjen, 2017). A 20-kb and 40-kb PacBio SMRTbell library was prepared by BGI (https://www.bgi.com) with 20 and 40-kb Blue Pippin size selection being performed prior to sequencing on a PacBio Sequel system (Pacific Biosciences, Menlo Park, CA).

Thirty-eight isolates of *P. pachyrhizi* (Supplementary Table 1) with known virulence on soybean genotypes containing the *Rpp1* and *Rpp5* resistance genes were sequenced using the DBN sequencing technology (BGI, https://www.bgi.com). Bad quality reads were trimmed using Trimmomatic version 0.36 to retain reads with a Phred value higher than 33.

### 2.3 Bioinformatic processes

#### 2.3.1 Genome assembling and reads mapping

FALCON and FALCON-Unzip algorithms (https://github.com/PacificBiosciences/FALCON/) were used to assemble the PacBio SMRT long-read sequencing data into highly accurate, contiguous, and correctly phased diploid genomes. We assessed the BUSCO scores after each assembly parameter adjustment to compare the improvement in the assemblies.

Minimap2 (Li, H., 2018) [parameters -ax] was used for DBN reads mapping against the assembled genomes. Purge Haplotigs (Roach et al., 2018) [parameters: default (also using the programs minimap2 and samtools faidx with default parameters)] were used to generate a haploid representation of the 71G2 genome (71G2h). RNA sequencing (RNA-seq) data combining *P. pachyrhizi* PPUFV02 transcripts from germinated spores and filtered from the soybean-*P. pachyrhizi* interaction at 0, 12, 24, 48, 72, 96, and 168 hours post-inoculation was downloaded from the Mycocosm database (https://mycocosm.jgi.doe.gov). The RNA-seq data and DBN-seq reads were mapped to the 71G2h assembly using Minimap2 (Li, H., 2018) [parameters -ax].

#### 2.3.2 SNP polymorphism identification

Variant calling was performed using the Genome Analysis Toolkit (GATK) 4.2.0 [Parameters: gatk --java-options "-Xmx800g -XX:+UseParallelGC" HaplotypeCaller --sample-ploidy 2 --pcr-indel-model] (Van der Auwera & O'Connor, 2020) comparing the

genotype of the isolates of *P. pachyrhizi*, using their DBN-seq reads aligned on the 71G2h genome. The *Rpp1b* virulent isolates PHPA24 and PHPA26 (natural mutants from isolates PHPA25 and PHPA23, respectively obtained during the fungi multiplication) were compared with the avirulent isolates PHPA25 and PHPA23, and the isolate *Rpp5* virulent isolate PHPA12 was contrasted with the avirulent isolates PHPA11 and PHPA13 in pairwise comparisons (PHPA comparisons: 24x25; 26x23; 12x11; 12x13). Variations detected in both comparisons 24x25 and 26x23; and 12x11 and 12x13 were filtered retaining only the variations present in loci predicted to encode secreted genes. For that, the predicted secreted proteins were selected and the gene sequences encoding them in *P. pachyrhizi* isolates PPUFV02, K8108, and MT2006 were downloaded from the Mycocosm database. A manual screening was performed at the Mycocosm platform among the three genome assemblies available (PPUFV02, K8108, and MT2006) to eliminate sequences without expression data (RNA-seq coverage). Sequences from the *P. pachyrhizi* secreted proteins library obtained in previous studies (unpublished data) were included in these analyses. T-Coffee (Di Tommaso et al., 2011) was used to filter sequences with more than 90% of identity and the information obtained was organized into a *P. pachyrhizi* secretome database. This database was annotated using the National Center for Biotechnology Information (NCBI) reference sequences non-redundant proteins database (NR database v5) available on the NCBI website (https://www.ncbi.nlm.nih.gov/refseq/). All genes were also identified on 71G2h genome using the Blastn or Blastp software algorithms implemented for local analysis (e-value threshold $1^{-10}$), generating tabular and gff3 files (Camacho et al., 2009). Genotypic variations and polymorphisms among *P. pachyrhizi* isolates, previously identified using GATK (Van der Auwera & O'Connor, 2020) were filtered to maintain the variations of the secretome genes, which were manually checked using the Integrative Genome Viewer - IGV (Thorvaldsdóttir et al., 2013). The screening of regions that could be encoding candidate avirulence genes in contigs 29, 60, 72, 87, 183, 221, 254, and 325 of 71G2h genotype also was done in the IGV, using the 71G2 as reference genome and the gff3 file previously obtained, the mapped reads from RNA-seq and DBN-seq were used to support the analysis.

The nomenclature adopted for the candidate avirulence genes follows the candidate factor and its position: [*cAvr*(*1* or *5*)-*contig_position* in 71G2h]. In the case of genes with multiple copies also receive a gene name to identify the multigenic family: [*cAvr*(*1* or *5*)-*name-contig_position* in 71G2h].

Analyses using Interproscan (Jones et al., 2014), Depicter (Barik et al., 2020), and SignalP5 (Almagro et al., 2019) were performed in their web servers

(https://www.ebi.ac.uk/interpro/; http://biomine.cs.vcu.edu/servers/DEPICTER/ and https://services.healthtech.dtu.dk/services/SignalP-5.0/), respectively, to identify the known functional domains, disordered regions and the presence of signal peptides.

**2.4 Molecular identification and genotyping**

Sanger sequencing of the polymorphic region of the candidate *cAvr1-PHPA79-221_967Kb* was performed for 14 avirulent isolates and seven virulent isolates, and the genotyping of the candidate *cAvr5-138_913Kb* was done with two virulent and two avirulent isolates. Polymerase Chain reactions (PCR) were performed with the Hotfire DNA polymerase (Solis Biodyne) using the oligonucleotides PHPA79_CdS-35F (CCTATTGAATCTAGTTTGGTGTCG) and PHPA79_CdS-35R (AATTACCAGGTTAGCCTTTTCTTG) with melting temperature of 59ºC for the gene candidate *cAvr1-PHPA79-221_967Kb*, and oligonucleotides cAVR5-F1 (CCCTTGAACTACGTTTGAGTCCAGCGGT) and cAVR5-R6 (GATGCAAGGCTAAAACTCTGGCGCGTCA) with melting temperature of 67ºC for the candidate *cAvr5-138_913Kb*, following instructions of the polymerase manufactures. The PCR products were analyzed by electrophoresis in agarose gel (1%), purified with BigDye XTerminator Purification Kit (Thermo Fisher), and sequenced at SeqStudio Genetic Analyzer (Thermo Fisher) using the oligonucleotides PHPA79_CdS-35F and cAVR5-F3 (TTCGGGGACTCGGTGATGGGGCATA).

**2.4.1 Race amplification and sequencing of cAvr5 gene**

Total RNA was isolated from germinating spores of isolate PHPA 12 and PHPA 13 using the RNeasy plant mini kit (Qiagen) and treated with DNAse (DNase Max, Qiagen). This treated total RNA was used for the Rapid Amplification of Complementary DNA Ends (RACE) (SMART RACE cDNA Amplification Kit, Clontech - Takara). The RACE cDNA products were used for PCR using the Hotfire DNA polymerase and the UPM oligonucleotide (SMART RACE kit) combined with gene-specific primers cAVR5-F1, cAVR5-F3, cAVR5-F6 (TGACGCGCCAGAGTTTTAGCCTTGCATC), cAVR5-F8 (TCTACGAGAGCTGCCTTTGGTTCAGA), cAVR5-R3 (TATGCCCCATCACCGAGTCCCCGAA) and cAVR5-R6, with the melting temperature of 68ºC. The PCR products analysis, purification, and sequencing were done as related before, but using the oligonucleotides cAVR5-F3, cAVR5-R3, cAVR5-F6, cAVR5-R6, and

cAVR5-F8. The ORFs were predicted with ORF Finder at the NCBI website. All the procedures using commercial kits were performed following the manufacturer's instructions without adaptations.

**2.5 Transient expression studies in *Nicotiana benthamiana***

The predicted alleles of *cAvr1* genes without introns (*cAvr1-PHPA79-221_967Kb*, *cAvr1-PHPA79-254_1125Kb,* and *cAvr1-PHPA79-254_1151Kb*) were synthesized by Epoch life science (epochlifescience.com) and cloned in plasmid pBluescript SK (-). The synthetic genes were used as templates for gene amplification by PCR using specific oligonucleotides. The alleles of gene *cAvr1-PHPA79-221_967Kb* were amplified with oligonucleotides PHPA79-F1 (CACCATGTATGGCTTTTTACCAAGCG) and PHPA79-R1 (CTAGTTTGGTGTCGCTGTTACAG) and the gene *cAvr1-PHPA79-254_1125Kb* and the alleles of gene *cAvr1-PHPA79-254_1151Kb* were amplified with PHPA79-F2 (CACCATGTTTGTACCGAGCGAACCATC) and PHPA79-R2 (CTAGTTTGGCGTTGCCGATACAGTG). For all amplification, the genic region encoding the predicted secretion signal peptide was excluded, and in the case of the sequences used for confocal microscopy in assays with the GFP cloned at the N-terminal extremity, oligonucleotides with the same sequence of PHPA79-R1 and PHPA79-R2 without the three first nucleotides (CTA, that correspond to the stop codon in the reverse sequence) were used. Both alleles of the gene *cAvr5-138_913Kb* were amplified from the cDNA of isolate PHPA 13 obtained from the RACE using the oligonucleotides cAVR5-F (CACCTTCAAAACAGAGCTCCTG) and cAVR5-R (TTAGCTCATTTTTTTAAGTG). All PCRs were performed with Hotfire DNA polymerase, according to the manufacturer's recommendations and specific melting temperatures of 61ºC in the amplification of gene *cAvr5-138_913Kb* and 58ºC in the other amplifications. Electrophoresis and sequencing were performed as previously described, using the specific oligonucleotides for each sequence.

**2.5.1 Subcellular localization assay**

The gene *cAvr1-PHPA79-254_1125Kb* and the alleles of genes *cAvr1-PHPA79-221_967Kb* and *cAvr1-PHPA79-254_1151Kb* without the sequence encoding the secretion signal peptide were cloned into pK7FWG2 and pK7WGF2 expression vectors and transformed into *Agrobacterium tumefaciens* GV3101 as described by Höfgen and Willmitzer (1988). *A. tumefaciens* clones with pMP90::*Aty-TIP-mCherry* (Saito et al., 2002)

and pMP90::*WWP1-mCherry* (Calil et al., 2018) were used in the transient transformation as vacuole and nucleus compartment markers, respectively. An isolated colony from each transformant was grown in Luria-Bertani (LB) medium (Sambrook et al., 1989) with corresponding antibiotics [rifampicin (100 μg $ml^{-1}$) and ampicillin (50 μg $ml^{-1}$) for pMP90::*WWP1-mCherry* selection; rifampicin (100 μg $ml^{-1}$), spectinomycin (30 μg $ml^{-1}$) and gentamicin (25 μg $ml^{-1}$) for pMP90::*AtPIP2A-mCherry*, and transformants in pK7FWG2 and pK7WG2 plasmids] were grown at 28°C in a shaking incubator at 180 rpm for 16 hours. Cultures were centrifuged and the pellets were resuspended and adjusted in different mixtures containing one candidate allele protein and one of the two cellular compartments markers to a final $OD_{600}$ 0.2 using the infiltration solution [10 mM MES, 10 mM $MgCl_2$, 200 μM acetosyringone (pH 5.6)]. *Agrobacterium* suspensions were incubated at room temperature for three hours and then used to infiltrate *N. benthamiana* leaves with a needleless syringe. Each inoculation replicate had one candidate gene and one cellular compartment marker. Three independent plants were transformed with *A. tumefaciens* and kept at 22°C for 72 hours and afterward analyzed in a laser confocal scanning microscope (LSM 510 META, Carl Zeiss, Oberkochen, Germany). *A. tumefaciens* cloned with the empty vectors were used as the control for GFP expression. GFP was excited at 488 nm and then captured at 550 nm. mCherry was excited at 554 nm and captured at 560-615nm. The images were processed with ZEN2.3 software (Carl Zeiss).

### 2.5.2 ETI and PTI suppression assays

The alleles of gene candidates *cAvr5-138_913Kb*, *cAvr1-PHPA79-221_967Kb*, *cAvr1-PHPA79-254_1125Kb,* and *cAvr1-PHPA79-254_1151Kb*, without the sequence encoding the secretion signal peptide, were cloned into pEDV6 (effector detector vector) (Sohn et al., 2007), transformed into *Escherichia coli* TOP10 (Thermo Fisher) and then, transformed into *Pseudomonas fluorescens* EtHAn (Effector-to-Host-Analyzer) - PfE, with functional Type Three Secretion System (TTSS)(Thomas et al., 2009) by standard triparental mating using *E. coli* HB101 (pRK2013) as helper strain (Fabro et al., 2011; Badel et al., 2013). *E. coli* cells were grown in LB medium containing kanamycin (50 μg $ml^{-1}$) at 37°C and *Pseudomonas* cells were grown in King's B (KB) medium (King et al., 1954) at 28°C with corresponding antibiotics [PfE: chloramphenicol (30 μg $ml^{-1}$); PfE::pEDV6: chloramphenicol (30 μg $ml^{-1}$) and gentamicin (25 μg $ml^{-1}$)]. *Pseudomonas syringae* pv. *garcae*

(Psg), the plant pathogenic bacteria used to induce HR response in PTI assay, was also cultivated in KB medium with rifampicin (100 μg $ml^{-1}$).

An isolated colony from each transformant was grown at 28°C in a shaking incubator at 180 rpm for approximately 16 hours. Cultures were centrifuged and resuspended in 10 mM $MgCl_2$. For PTI assays, the $OD_{600}$ was adjusted to 0.2 for PfE::cAVR (PfE transformed with the candidate avirulence genes) or PfE::Ø (empty vector) and 0.025 for Psg. Suspensions of PfE were infiltrated in 30 independent *N. benthamiana* leaves and after seven hours Psg cultures were also infiltrated next to the first inoculation, creating an intersection area between infiltrated regions. PfE::*Ec23*, a candidate effector from *P. pachyrhizi* described as PTI/ETI suppressor (Qi et al., 2016), was used as a positive control. The tissue in the overlapping area was evaluated 48 hours after inoculation, the presence of HR in the co-infiltrated area was an indication of PTI suppression. Inoculations only were evaluated in leaves where the HR was not observed in the area inoculated with the negative control (PfE::Ø).

For ETI assays, *N. benthamiana* plants were infiltrated with the same PfE suspensions, but in this case, the bacteria were diluted in 10 mM $MgCl_2$ for $OD_{600}$ 0.15 and mixed with PfE transformed with pVSP6::*avrB*, for *avrB* gene expression, at $OD_{600}$= 0.1, previously grown in LB medium with chloramphenicol (30 μg $ml^{-1}$) e kanamycin (50 μg $ml^{-1}$) and co-inoculated. *N. benthamiana* contains endogens NLR genes which recognize the *AvrB* effector (Kessens et al., 2014), therefore when Pf EtHAn translocates AvrB into *N. benthamiana* cells, by plasmid pVSP61::*AvrB*, it induces an HR AvrB-dependent (ETI response), whereas Pf EtHAn expressing pEDV6 empty did not induce HR (Innes et al., 1993). The non-HR induction in the co-infiltration region between Pf EtHAn expressing AvrB and Pf EtHAn expressing Avr genes candidates is an indication of ETI suppression. PfE::*Ec23* was also used as a positive control (ETI suppressor). Mixes with both bacteria were used to infiltrate 30 independent leaves. Inoculations only were evaluated in leaves where the HR was not suppressed by the negative control (PfE::Ø).

### 2.5.3 Coexpression of *Rpp5* and cognate *candidate avirulence 5* gene

The *Rpp5* resistance gene and one gene linked to the *Rpp5*, denominated *7900*, are associated with soybean resistance against *P. pachyrhizi* (unpublished data). The genes *Rpp5* and *7900* were synthesized using chemistry synthesis by Epoch life science (epochlifescience.com). Both genes were synthesized without introns sequence under the

control of the soybean (*Glycine max*) ubiquitin promoter (GmubiXL) (and nopaline synthase terminator (Tnos) from *Agrobacterium tumefaciens* in pBluescript KS (-) plasmid. The *Rpp5* constructions (GmubiXL_Rpp5N and GmubiXL_Rpp5M) were cloned in plasmid pGlymax-Bar-MR8 also synthesized by Epoch life science (epochlifescience.com) (Figure 1) and the gene *cAvr5-138_913Kb* were cloned into pCambia3301 using restrictions enzymes and they were transformed separately into *A. tumefaciens* GV3101 as described by Höfgen and Willmitzer (1988). Bacteria multiplication and purification were performed as previously described for *A. tumefaciens* clones for the subcellular localization assay, using the corresponding antibiotics [kanamycin (50 µg $ml^{-1}$) for pCambia3301 and spectinomycin (30 µg $ml^{-1}$) for pGmM8]. The following suspension treatments were prepared: one with two bacteria clones in the suspensions, mixing the clones harboring genes *Rpp5* and *cAvr5-138_913Kb*; a second treatment mixing the clones harboring genes *Rpp5, cAvr5-138_913Kb* and *7900*; negative control treatments with bacteria clones alone, and a clone with pCambia3301:*bar* for transient transformation and gene expression control. The bacterial suspensions were adjusted to $OD_{600}$ 0.2 using the infiltration solution previously described for *Agrobacterium* infiltration and the *N. benthamiana* leaves were infiltrated with a needless syringe. Evaluation of induced HR was done 48 hours after the infiltration.

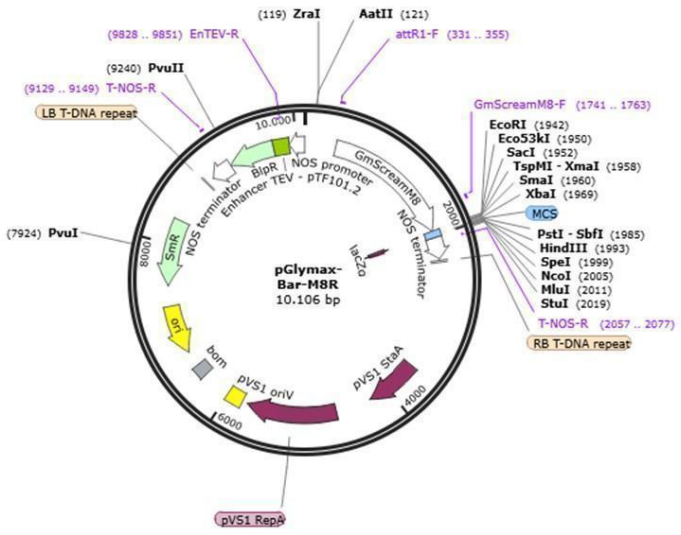



**Figure 1:** pGlymax-Bar-M8R plasmid with the phosphinothricin acetyltransferase (Bar) reporter gene and multiple cloning site (MSC) under control of GmscreamM8 promoter from soybean (*Glycine max*), designed by chemistry synthesis.

# 3. RESULTS

## 3.1 Genome assemblies of *P. pachyrhizi* isolate 71G2

Long-read sequencing data were used to assemble the genome of *P. pachyrhizi* isolate 71G2. The 71G2 diploid genome assembly (71G2d) has 1.411 Mbp and the genomes' completeness was accessed by BUSCO using basidiomycetes database, resulting in more than 90% of complete sequences (Table 1). The metrics obtained for genomes assemblies are summarized in Table 1.

In addition to the 71G2 diploid assembly with DNA from both nuclei, a second assembly for 71G2 isolate using purging haplotigs (Roach et al., 2018) was performed to generate a haploid representation of the 71G2 genome (71G2h). This haploid genome was used to align DBN sequencing reads from other *P. pachyrhizi* isolates containing alleles from both nuclei in the haploid assembly representation. The genome completeness of the 71G2h isolate was accessed by BUSCO and a decrease in the duplicated gene content (from 84,1% to 21,8%), an increase in single-copy content (from 6,3% to 67,2%), a small loss of genes (missing genes, from 8,4% to 9,8%) and consequent reduction on total completeness to 89,0% were observed (Table 1).

**Table 1**: Genome assembly and BUSCO summary result for *P. pachyrhizi* isolate 71G2 diploid (71G2d) and representative haploid (71G2h) genomes.

| **Genome assemblies** | | |
|---|---|---|
| | **Assembly** | |
| | **71G2d[a]** | **71G2h[b]** |
| Contig number | 7.420 | 1385 |
| Max contig length (bp) | 5.612.061 | 5.612.061 |
| N50 length (bp) | 884.602 | 1.357.846 |
| L50 | 410 | 227 |
| Total length (Mbp) | 1.411,57 | 1.009,95 |
| GC% | 37,30 | 37,35 |
| **Completeness of the genome assemblies (BUSCO)** | | |
| | **Assembly** | |
| | **71G2d[a]** | **71G2h[b]** |
| Complete total (S + D) | 1.594 (90,4%) | 1.569 (89,0%) |
| Complete and single-copy (S) | 111 (6,3%) | 1185 (67,2%) |
| Complete and duplicated (D) | 1.483 (84,1%) | 384 (21,8%) |
| Fragmented | 21 (1,2%) | 21 (1,2%) |
| Missing | 149 (8,4%) | 174 (9,8%) |
| Total | 1.764 (100%) | 1.764 (100%) |

[a]d: diploid; [b]h: haploid

The DBN sequencing from the whole genome of the 71G2 and other 37 *P. pachyrhizi* isolates were trimmed and aligned against the 71G2h and visualized using IGV program. The presence of polymorphic sites in some genotypes indicates at least one variant allele in the diploid data mapped in the haploid representation, while regions without polymorphisms indicate homozygotic loci.

### 3.2 Identification of candidate avirulence loci of *P. pachyrhizi* by comparative genomics

A database of *P. pachyrhizi* genes from isolates K8108, MT2006, and PPUFV02 that were predicted to encode secreted proteins was built for local analyses. The database was also

implemented with genes from protein libraries obtained in previous studies (unpublished data) and proteins predicted by Link et al. (2014) and Kunjeti et al. (2016). All these sequences were manually checked in genome assemblies at the Mycocosm platform to maintain only genes with RNA coverage. Then, a curated database was generated from the predicted secreted proteins of all isolates. By eliminating repeated sequences, a database of unique 785 sequences of predicted proteins was obtained. This database designated "*P. pachyrhizi* secretome" was also annotated using the NCBI RefSeq non-redundant proteins database (Supplementary data). From the 785 predicted proteins, 417 proteins (53,1%) showed similarity with proteins already annotated in other species, being most of them annotated as hypothetical proteins of rust fungi, and 368 proteins (46,9%) did not show significant similarity against other proteins of the NCBI database. The secretome's database was used as a query in local analysis using Blastn against 71G2h assembly. Some genes were aligned in more than one loci in the genome, resulting in a total of 1516 mapped loci.

In previous studies, eleven isolates (PHPA 01, 03, 05, 06, 09, 22, 24, 26, 30, 31, and 71G3) were identified as virulent in soybean genotypes containing *Rpp1b* resistance gene, and two isolates (PHPA 12 and 28) virulent on soybean genotypes with *Rpp5* (unpublished data). The DBN-seq data of the 38 *P. pachyrhizi* isolates (including the 71G2 diploid data) were individually mapped against 71G2h and genotype variations were observed among the *P. pachyrhizi* isolates. The Genome analysis toolkit (GATK) was used to identify polymorphism in isolates with differences in the virulence/avirulence phenotype. Two pairwise comparisons between virulent/avirulent isolates for each target avirulence gene were performed, then variant call positions were filtered only for variations present in loci containing genes represented in the secretome database. The regions or genes that could be associated with the virulence against soybean genotypes carrying *Rpp1b* or *Rpp5*, were identified as candidate avirulence genes (candidates *Avr1* or *Avr5*). In the case of the candidates *Avr5*, single nucleotide variations (SNPs) were identified in three different genes and 16 SNPs at one transcribed region, possibly spanning two ORFs, resulting in four regions possibly encoding the Avr5 (Figure 5). The candidates *Avr* were identified, according to their loci, as *cAvr5-47_1128Kb* (Figure 2), *cAvr5-134_549Kb* (Figure 3), *cAvr5-310_66Kb* (Figure 4), and *cAvr5-138_913Kb* (Figure 5). In the case of *cAvr1*, it was not possible to identify *Avr1* candidates in secretome loci using GATK tools.

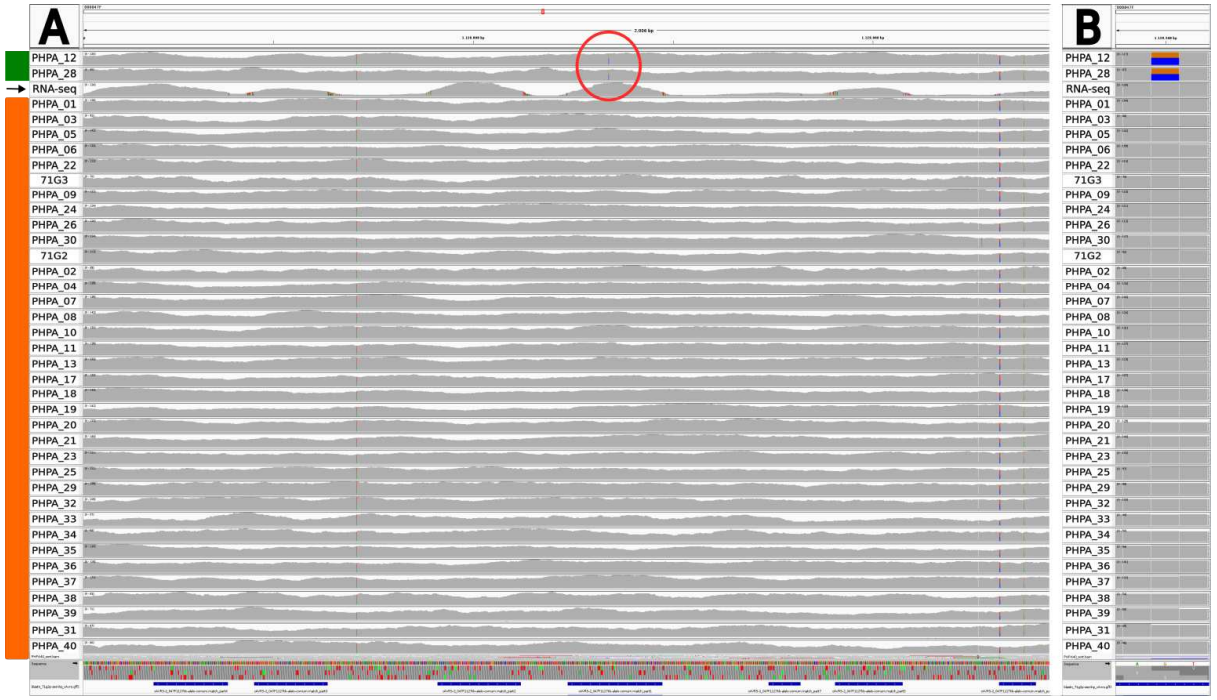



**Figure 2:** Locus of candidate *cAVR5-47_1128Kb* viewed in IGV program (adapted). Two virulent *P. pachyrhizi* isolates (green bar) were compared with 35 avirulent ones (orange bar). The entire gene is shown in A, where the blue lines at the bottom of the figure indicate the predicted exons of the gene and the red circle emphasizes the polymorphic position among the contrasting isolates that is highlighted in B. The third line (indicated with a black arrow), separating the genome of virulent and avirulent isolates, is the RNA-seq coverage. Vertical lines represent the nucleotides: grey when it is identical to the reference genome and colored (Green: A; Blue: C; Orange: G; Red: T) when the nucleotide is different from the reference.

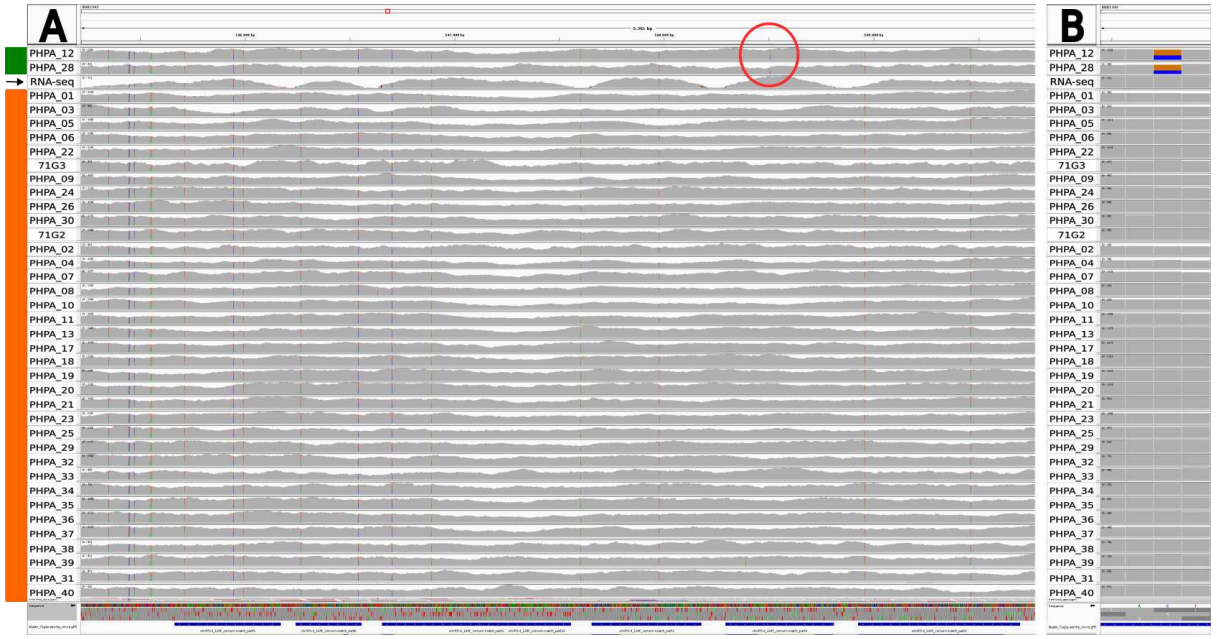



**Figure 3:** Locus of candidate *cAVR5-134_549Kb* viewed in IGV program (adapted). Two virulent *P. pachyrhizi* isolates (green bar) were compared with 35 avirulent ones (orange bar). The entire gene is shown in A, where the blue lines at the bottom of the figure indicate the predicted exons of the gene and the red circle emphasizes the polymorphic position among the contrasting isolates that is highlighted in B. The third line (indicated with a black arrow), separating the genome of virulent and

avirulent isolates, is the RNA-seq coverage. Vertical lines represent the nucleotides: grey when it is identical to the reference genome and colored (Green: A; Blue: C; Orange: G; Red: T) when the nucleotide is different from the reference.

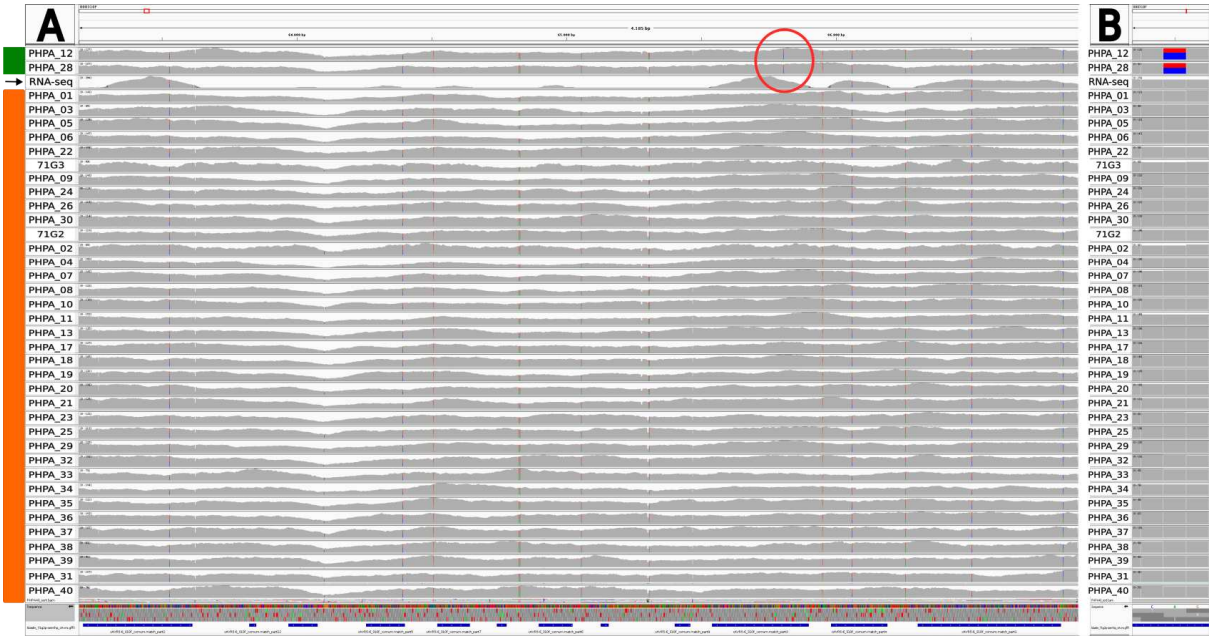



**Figure 4:** Locus of candidate cAVR5-310_66Kb viewed in IGV program (adapted). Two virulent *P. pachyrhizi* isolates (green bar) were compared with 35 avirulent ones (orange bar). The entire gene is shown in A, where the blue lines at the bottom of the figure indicate the predicted exons of the gene and the red circle emphasizes the polymorphic position among the contrasting isolates that is highlighted in B. The third line (indicated with a black arrow), separating the genome of virulent and avirulent isolates, is the RNA-seq coverage. Vertical lines represent the nucleotides: grey when it is identical to the reference genome and colored (Green: A; Blue: C; Orange: G; Red: T) when the nucleotide is different from the reference.

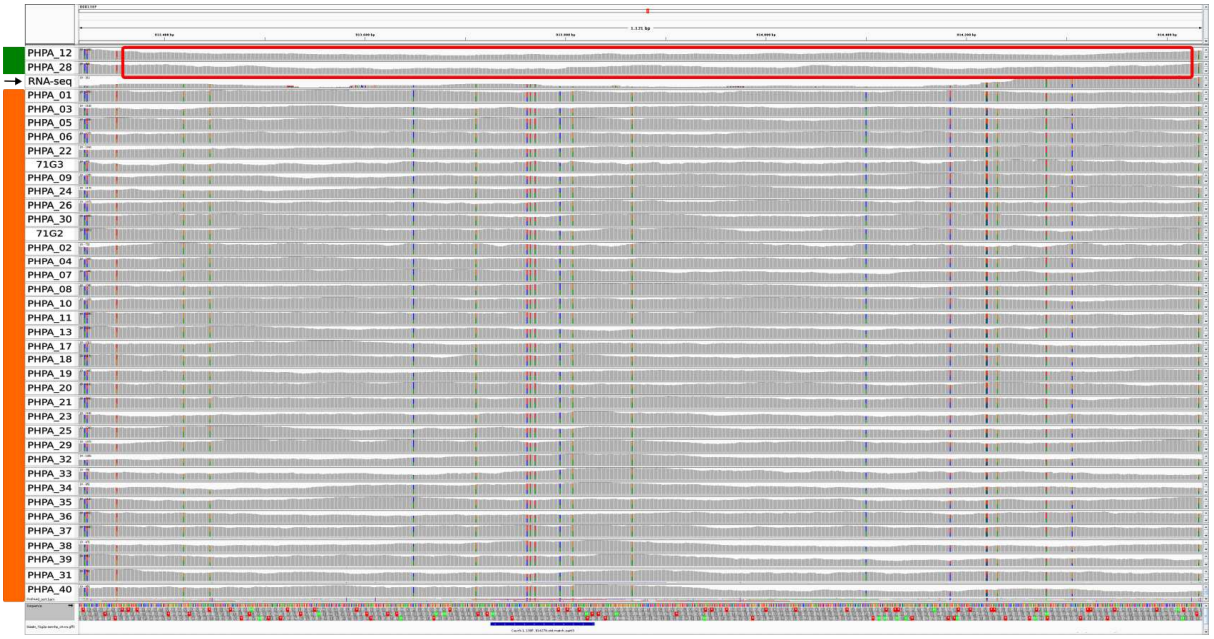



**Figure 5:** Locus of candidate *cAvr5-138_913Kb* viewed in IGV program (adapted). Two virulent *P. pachyrhizi* isolates (PHPA 12 and 28, green bar) were compared with 35 avirulent ones (orange bar). The entire region with loss of heterozygosity in the genome of the isolates PHPA 12 and 28 is

highlighted in the red rectangle. The third line (indicated with a black arrow), separating the genome of virulent and avirulent isolates, is the RNA-seq coverage and the blue horizontal lines at the bottom represent the *cAvr5-138_913Kb* gene prediction. Vertical lines represent the nucleotides: grey when it is identical to the reference genome and colored (Green: A; Blue: C; Orange: G; Red: T) when the nucleotide is different from the reference.

Each locus of 71G2h secretome was also manually analyzed to identify genic and genotypic variations among *P. pachyrhizi* isolates that could be associated with the phenotype of virulence or avirulence in soybean genotypes containing *Rpp1b* since no variations were identified using GATK. From the 1516 loci resulting from the secretome alignment against 71G2h, 750 were heterozygous, while 766 were homozygous at the predicted coding region. The screening performed for candidate *Avr1* made it possible to observe that five of the eleven virulent isolates (PHPA 01, 03, 05, 06, and 22) on genotypes containing the *Rpp1b* gene candidate had genomic variations in a large region, affecting at least eight large contigs (Contigs 29, 60, 72, 87, 183, 221, 254 and 325, on 71G2h assembly). These five isolates are homozygous in this region, while the other six isolates are heterozygous (Figure 6). Since the loss of heterozygosity, if affecting the avirulence gene (*Avr*), could be related to the virulence of five of the eleven virulent isolates, another SNP screening was performed in these eight contigs (29, 60, 72, 87, 183, 221, 254, and 325 on 71G2h *P. pachyrhizi* genotype). The search aimed to find genomic variations that could be associated the virulence phenotype in the six other virulent isolates that did not present the loss of heterozygosity genotype. With these new searches, three loci variating (*cAvr1-60_267Kb*, *cAvr1-221_977Kb,* and the *cAvr1-PHPA79-221_967Kb*) (Figures 7, 8, and 9) among virulent and avirulent isolates were identified. The locus *cAvr1-PHPA79-221_967Kb* also was present in the secretome database and identified as *PHPA_79*.

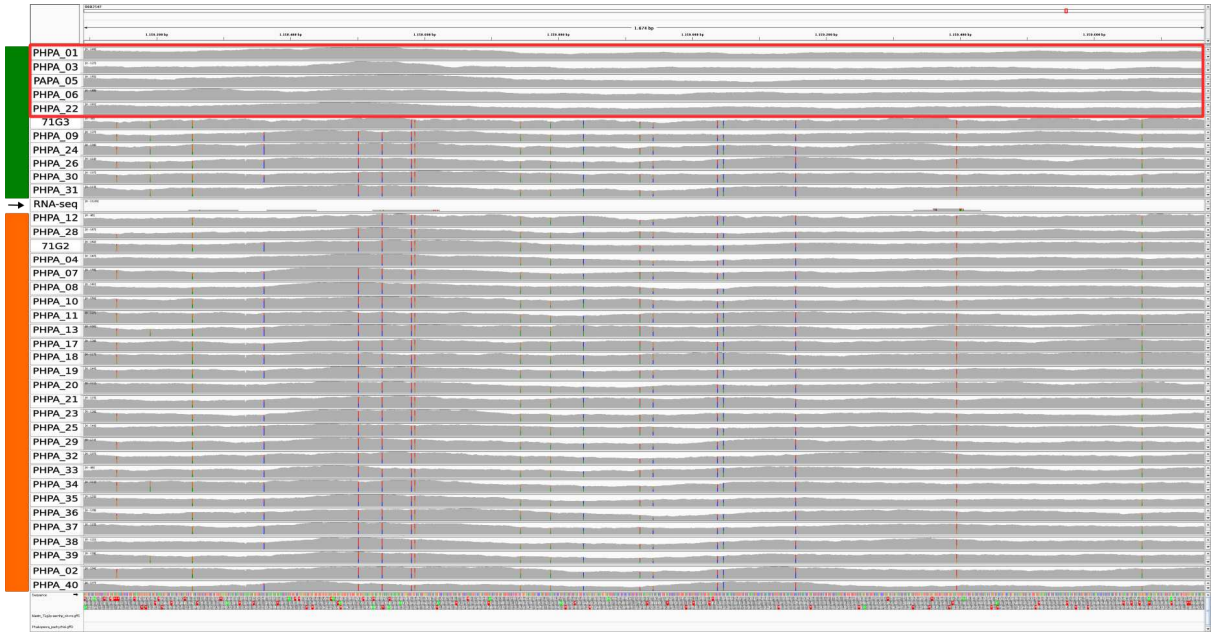



**Figure 6:** Loss of heterozygosity in contig 254 of the *P. pachyrhizi* isolates PHPA 01, 03, 05, 06, and 22 compared with heterozygous isolates viewed in the IGV program (adapted). The green bar indicates isolates virulent on soybean genotypes containing *Rpp1b* and the orange bar indicates the avirulent ones. The red box highlights the loss of virulence region in the genome of the five homozygotic isolates. The 12th line (indicated with a black arrow), separating the genome of virulent and avirulent isolates, is the RNA-seq coverage. Vertical lines represent the nucleotides: grey when it is identical to the reference genome and colored (Green: A; Blue: C; Orange: G; Red: T) when the nucleotide is different from the reference.

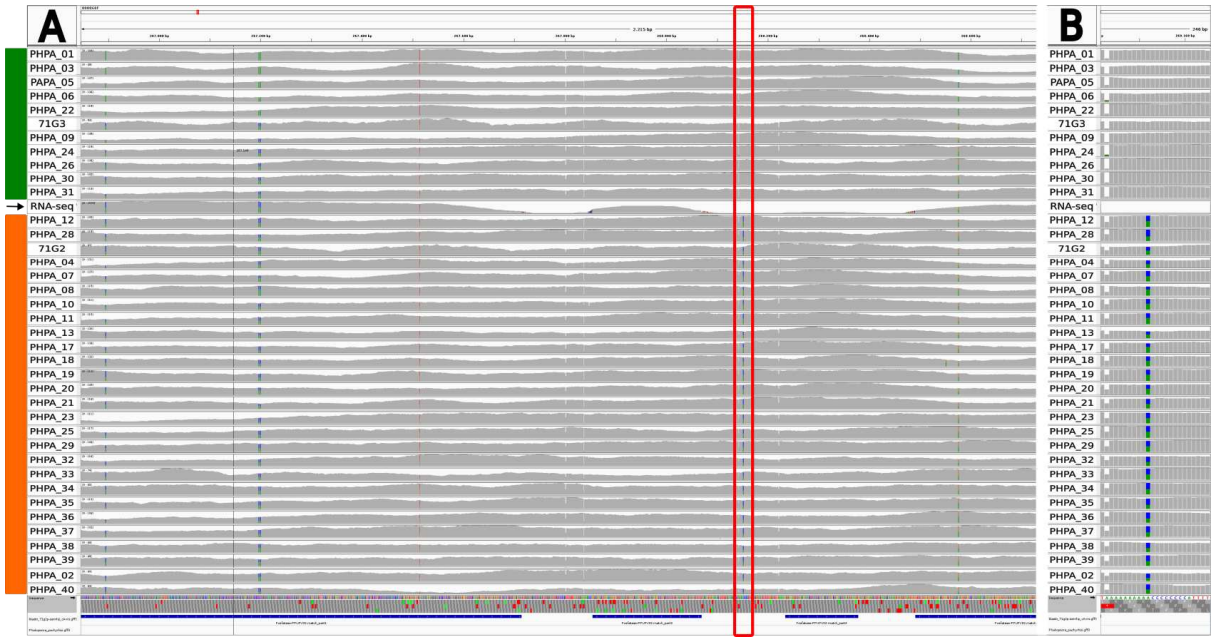



**Figure 7:** Locus of candidate *cAvr1-60_267Kb* viewed in IGV program (adapted). Eleven virulent *P. pachyrhizi* isolates (green bar) were compared with 26 avirulent ones (orange bar). Part of the predicted gene is shown in A, where the blue lines at the bottom of the figure indicate the predicted exons of the gene and the red circle emphasizes the polymorphic position among the contrasting isolates that is highlighted in B. The 12th line (indicated with a black arrow), separating the genome of virulent and avirulent isolates, is the RNA-seq coverage. Vertical lines represent the nucleotides: grey

when it is identical to the reference genome and colored (Green: A; Blue: C; Orange: G; Red: T) when the nucleotide is different from the reference.

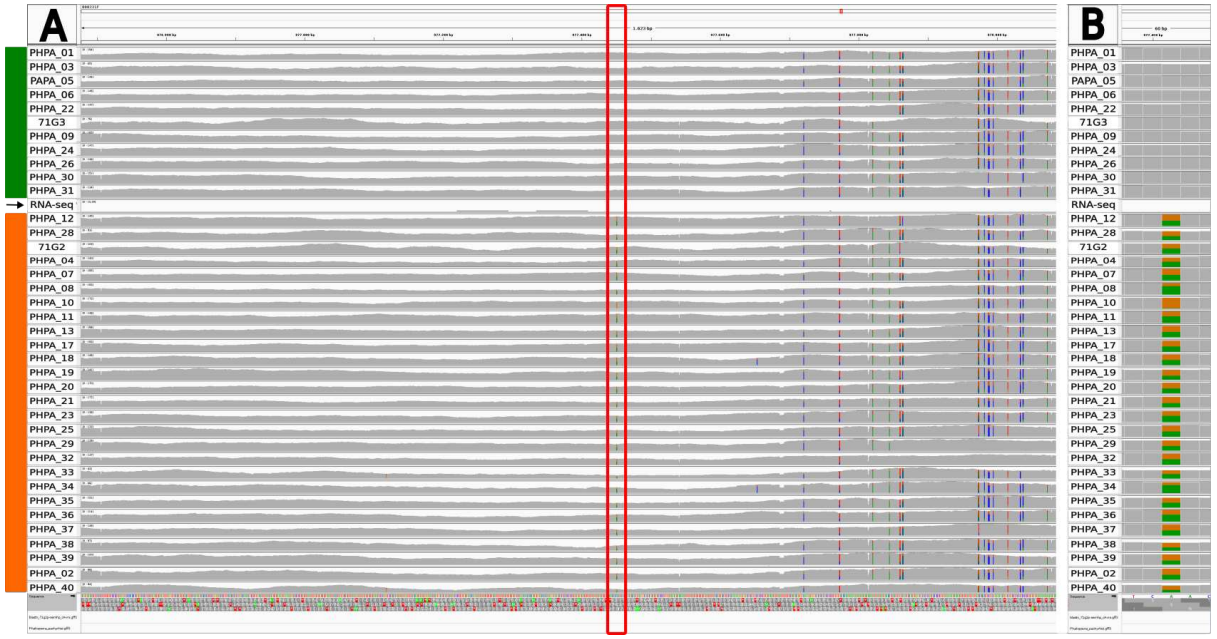



**Figure 8:** Locus of candidate *cAvr1-221_977Kb* viewed in IGV program (adapted). Eleven virulent *P. pachyrhizi* isolates (green bar) were compared with 26 avirulent ones (orange bar). Part of the predicted gene is shown in A, where the blue lines at the bottom of the figure indicate the predicted exons of the gene and the red circle emphasizes the polymorphic position among the contrasting isolates that is highlighted in B. The 12th line (indicated with a black arrow), separating the genome of virulent and avirulent isolates, is the RNA-seq coverage. Vertical lines represent the nucleotides: grey when it is identical to the reference genome and colored (Green: A; Blue: C; Orange: G; Red: T) when the nucleotide is different from the reference

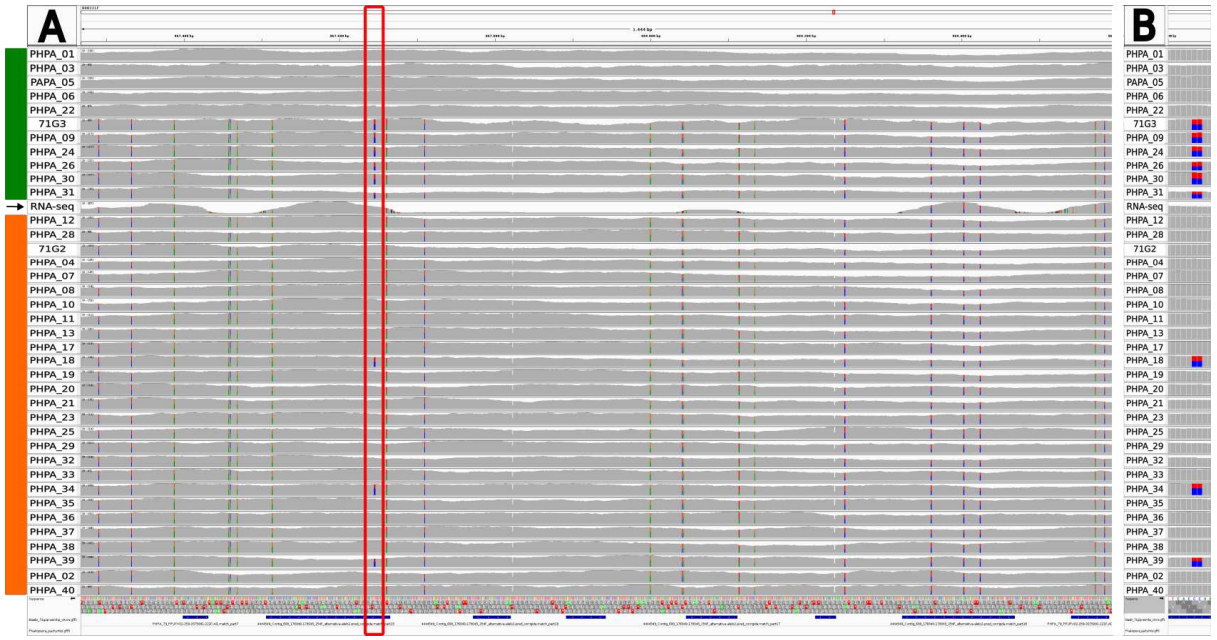



**Figure 9:** Locus of candidate *cAvr1-PHPA79-221_967Kb* viewed in IGV program (adapted). Eleven virulent *P. pachyrhizi* isolates (green bar) were compared with 26 avirulent ones (orange bar). Part of

the predicted gene is shown in A, where the blue lines at the bottom of the figure indicate the predicted exons of the gene and the red circle emphasizes the polymorphic position among the contrasting isolates that is highlighted in B. The 12th line (indicated with a black arrow), separating the genome of virulent and avirulent isolates, is the RNA-seq coverage. Vertical lines represent the nucleotides: grey when it is identical to the reference genome and colored (Green: A; Blue: C; Orange: G; Red: T) when the nucleotide is different from the reference

### 3.3 *In silico* characterization of the identified candidate avirulence genes

#### 3.3.1 Candidate avirulence genes of *P. pachyrhizi* recognized by *Rpp5* (c*Avr5*)

Four loci were identified as associated with genes that could encode Avr5. The candidate genes *cAvr5-47_1128Kb*, *cAvr5-134_549Kb,* and *cAvr5-310_66Kb* were identified due to a single nucleotide variation (Figures 2-4) and the candidate *cAvr5-138_913Kb* was identified due to 16 SNPs (Figure 5) in an expressed region where two possible genes could be affected. In the case of the *cAvr5-138_913Kb*, the genotypic variation occurred in a region of loss of heterozygosity with approximately 1 Kb where the two virulent isolates are homozygous and the avirulent isolates are heterozygous. On the other hand, the three other candidates had the genotypic variation in the opposite direction: for the variant nucleotide, all avirulent isolates are homozygous, while the two virulent isolates are heterozygous. All these three candidates were predicted as proteins at the reference genome assemblies available at Mycocosm and their expression is supported by RNA-seq transcripts (Figures 2-5).

The candidate *cAvr5-47_1128Kb* was identified in the contig 47F of the 71G2h genome, encoding one protein with 390 amino acids and two allelic variants. The virulent isolates PHPA 12 and 28 showed a nucleotide change (C to G), not observed in the avirulent isolates (Figure 2), that resulted in an amino acid substitution (S225T). This candidate avirulence gene encodes the protein ID 3726186 in the reference genome of PPUFV02 at Mycocosm database annotated as a hypothetical protein present in rust fungi (with 35% similarity). Interproscan predictions detected a signal peptide cleavage consensus site between the 26th and 27th amino acids in the N-terminal end and a transmembrane domain between the 148th and 170th amino acids (Table 2).

The candidate *cAvr5-134_549Kb*, identified in the contig 134F of the 71G2h genome, encodes a 355 amino acids protein and possesses two allelic variants. The isolates PHPA 12 and 28 exhibited a mutation changing one nucleotide (C to G) that resulted in a premature stop codon after the 284th amino acid (Figure 3). This protein was not predicted at the reference genome PPUFV02, but it was annotated in the genome assembly of the isolate MT2006 as the protein ID 6548230. This protein has no significant similarity to any other

annotated protein, being present only in *P. pachyrhizi*. Interproscan predicted a membrane-bound domain along the protein, except for the 36 first amino acids that were predicted to be the signal peptide region also confirmed by the Protter tool (Table 2).

The candidate *cAvr5-310_66Kb*, identified in the contig 310F of the 71G2h genome, encodes a protein with 719 amino acids and it also has two allelic variants. The nucleotide changes that occurred in isolates PHPA 12 and 28 (A to G) (Figure 4) resulted in an amino acid substitution (T278A). This avirulence candidate gene encodes the protein ID 692760 in the reference genome of PPUFV02, annotated as an alpha/beta hydrolase (42% similarity) present in several basidiomycetes. Interproscan prediction detected a signal peptide cleaved site between the 25th and 26th amino acids and a serine carboxypeptidase domain between the 340-685th amino acids (Table 2).

The *cAvr5-138_913Kb* was identified in a region where RNA-seq transcripts support the expression of two genomic regions. Different gene predictions are possible for this region in the genome assemblies available at Mycocosm. To determine the extension of the transcribed region, the Rapid Amplification of Complementary DNA Ends (RACE) was performed and the amplicons were sequenced. Using the 5' RACE cDNA product, two regions were amplified using the universal oligonucleotide UPM combined with the sequence-specific oligonucleotides cAVR5-R3 for one segment and cAVR5-R9 for the second. For 3' terminal amplification, the 3' RACE was performed using cAVR5-F1 and cAVR5-F6. The oligonucleotides cAVR5-F3 and cAVR5-R3 were used for sequencing of the amplified fragment. A complete sequence of 615 nucleotides exhibiting the sequence of the oligonucleotide Smart IIA, inserted by the RACE adapter in the 5' region, and the poly-A segment in the 3' region was identified in the first amplified segment. For the second amplified segment, the oligonucleotides were used: cAVR5-F6, cAVR5-F8, and cAVR5-R6 for sequencing. It was recovered a larger fragment with approximately 1,2 Kb containing the oligonucleotide sequence of primer Smart IIA in the 5' region.

Using the NCBI ORF finder tool, the most probable ORFs spanning in each transcribed segment were identified. The first segment, obtained by sequencing with oligonucleotides cAVR5-F3 and R3, carried a possible coding region encoding a small peptide of 33 amino acids, and the second transcribed segment could encode a protein with 437 amino acids. The realignment of both predicted genes against the 71G2h genome demonstrated that no polymorphism among virulent and avirulent isolates was identified in the coding region that encoded the largest protein (with 437 amino acids). The coding region was also not affected by the loss of heterozygosity. However, the ORF encoding the smallest protein (with

33 amino acids) contained five polymorphic sites, that resulted in four amino acid changes in the predicted peptide. Therefore, the candidate *cAvr5-138_913Kb* was identified as the sequence encoding the small peptide with 33 amino acids. The small sequence of the candidate cAvr5-138_913Kb did not show similarity with any other peptide and known domain (Table 2).

To confirm the loss of heterozygosity in the region around the *cAvr5-138_913Kb,* the genomic region was amplified using the DNA from four isolates, two virulent and two avirulent isolates. The sequencing of amplified products corresponding to the candidate *cAvr5-138_913Kb* in the four isolates showed the same genotypic variation identified in the whole genome sequencing for isolates 12 and 28. In addition, it was demonstrated that two virulent isolates just have one allele of the gene, while the two avirulent isolates contain two alleles.

**Table 2:** Summarized results for *in silico* characterization of the candidate avirulence proteins.

| Analyzed gene/locus | Protein size (amino acids) | Annotation (Taxonomical specificity) | Signal peptide sequence | Cysteine residues | Predicted domains | Mutation[a] | Mutant isolates[a] |
|---|---|---|---|---|---|---|---|
| *cAvr5-47_1128Kb* | 390 | Hypothetical protein (Pucciniales) | Yes | 9 | Transmembrane | S225T | 12, and 28 |
| *cAvr5-134_549Kb* | 355 | No-hit (*Phakopsora* spp.) | Yes | 0 | Membrane-bound | Stop-codon | 12, and 28 |
| *cAvr5-310_66Kb* | 719 | Alpha/beta hydrolase (Fungi) | Yes | 6 | Serine carboxypeptidase | T278A | 12, and 28 |
| *cAvr5-138_913Kb* | 33 | No-hit (*Phakopsora* spp.) | No | 0 | None | Loss of heterozy-gosity | 12, and 28 |
| *cAvr1-60_267Kb* | 696 | Phosphatase 2C-like domain (Fungi) | No | 12 | PPM-type phosphatase-like | NA (Intron) | 01, 03, 05, 06, 09, 22, 24, 26, 30, and 31 |
| *cAvr1-221_977Kb* [b] | NA | NA | NA | NA | NA | NA | 01, 03, 05, 06, 09, 22, 24, 26, 30, and 31 |
| *cAvr1-PHPA79-221_967Kb* | 180 | No-hit (*Phakopsora* spp.) | Yes | 6 | Consensus disorder | Not affected [c] | 01, 03, 05, 06, and 22 |
| | 180 | No-hit (*Phakopsora* spp.) | Yes | 6 | Consensus disorder | G125S | 09, 17, 24, 26, 30, 31, 33, and 38 |
| *cAvr1-PHPA79-254_1125Kb* | 180 | No-hit (*Phakopsora* spp.) | Yes | 7 | Consensus disorder | Loss of heterozy-gosity | 01, 03, 05, 06, and 22 |
| *cAvr1-PHPA79-254_1151Kb* | 180 | No-hit (*Phakopsora* spp.) | Yes | 7 | Consensus disorder | Loss of heterozy-gosity | 01, 03, 05, 06, and 22 |

a: Mutation direction was supposed for the gain of virulence; b: There is not a prediction for the candidate *cAvr-221_977Kb* since its sequence was not determined; c: The nucleotide changes result in synonymous amino acid changes. NA = Not available.

### 3.3.2 Candidate avirulence genes of *P. pachyrhizi* recognized by *Rpp1b* (c*Avr1*)

Three regions that could contain *Avr1*, based on our previous screening, were analyzed. The candidate *cAvr1-PHPA79-221_967Kb* was previously obtained from a library of expressed segment tags (EST) denominated as *PHPA_79* (***Ph****akopsora* ***pa****chyrhizi*) (unpublished results) and was also included in the secretome database. The encoded protein of the gene *PHPA_79* was used as the query in tblastn searches against the genome of the *P.*

*pachyrhizi* isolate PPUFV02 at the Mycocosm website. Six predicted proteins related to the PHPA_79 were identified, three of them with more than 75% of similarity (PPUFV02 proteins ID: 3375891, 3684774, and 4444549) and three others more divergent, PPUFV02 proteins ID: 3254551, 3752077 and 3254573, with 62,4%, 59,32% and 56,9% of protein similarity, respectively. In *P. pachyrhizi* isolate 71G2h genome assembly, a sequence identical to the gene *PHPA_79* was identified on contig 221, which corresponds to the gene encoding the protein ID 3375891 from PPUFV02 isolate. In addition, two paralog genes located at contig 254F, separated for approximately 26 Kb**,** corresponding to the genes encoding the proteins ID 3684774 and 4444549, were also identified at the PPUFV02 genome assembly with a similar structure. In PPUFV02, the genes encoding the proteins ID 3254551 and 3254573 were located in the same contig, approximately 24 Kb apart, and the gene encoding the protein ID 3752077 was located in a different contig. In 71G2 genome assembly, three genes encoding these proteins (ID 3752077, 3254551, and 3254573) were localized in the same contig (092), two of them approximately 26 Kb apart and the third gene approximately 534 Kb distant from the closest paralog gene (Figure 10).

The genes of contig 092 encoding proteins 3752077, 3254551, and 3254573 were not polymorphic between isolates and could not explain the phenotypic variation observed among virulent and avirulent isolates, therefore they were not considered as the candidates for encoding the *Avr1*. The three other paralogs of *cAvr1-PHPA_79* (ID 3375891, 3684774, and 4444549) located in contigs 221 and 254F on the other hand were polymorphic, harboring mutations in the three paralog genes. Using the RNA-seq reads alignment and observing the most conserved fungi splicing site (5'GT, 3'AG), the prediction of genes encoding the proteins 3375891, 3684774, and 4444549 was edited for the most probable coding sequence, since small variations due to wrong predictions could change the reading frame and resulted in false amino acid variations. After the correction, the gene sequence encoding the protein 3375891 became identical to the *cAvr1-PHPA_79* previously obtained from an EST library (unpublished internal information), and proteins 3684774 and 4444549 became more similar to the protein encoded by *cAvr1-PHPA_79* (88,4% and 85,6% of protein similarity and 92,8% and 91,3% of gene identity, respectively).

The edited proteins received an "Ed-" prefix in their identification to differentiate them from the previously predicted form and the three paralog genes localized in the 71G2h genome were identified according to their position: *cAvr1-PHPA79-221_967Kb* encoding Ed-3375891, *cAvr1-PHPA79-254_1125Kb* encoding Ed-3684774 and *cAvr1-PHPA79-254_1151Kb* encoding Ed-4444549. The sequence of *PHPA_79* paralog genes

was not identified in other fungi, indicating that these genes are species-specific, at this moment, and they exhibited unknown functions. The encoded proteins from the three candidates *cAvr1-PHPA79* genes were predicted to be secreted, encoding a protein with 180 amino acids and six cysteine residues in the protein Ed-3375891 and seven in the proteins Ed-3684774 and Ed-4444549 (Table 2).

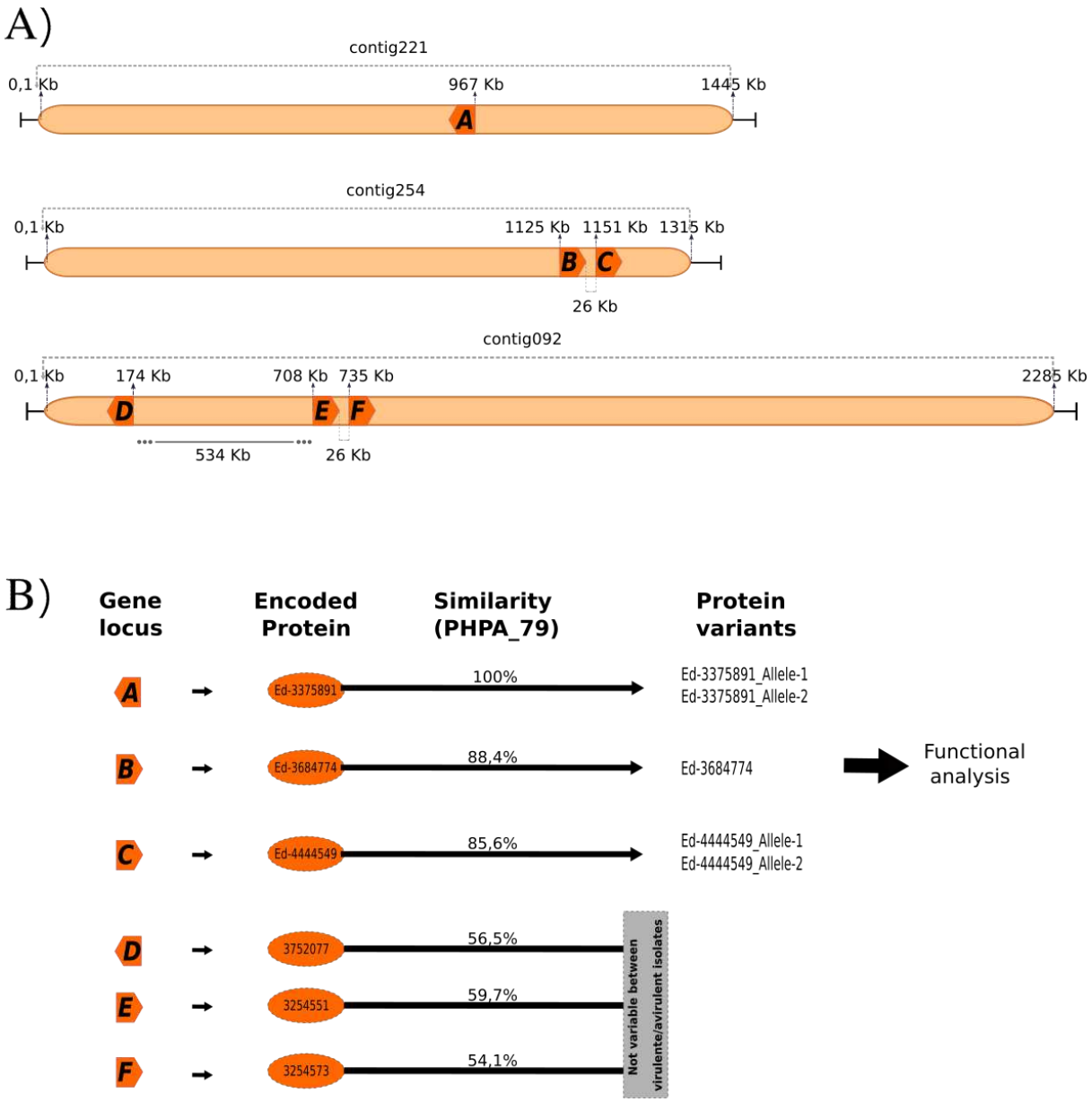



**Figure 10:** Genome structure of *PHPA_79* paralog genes (A) and identification of proteins and allelic variants. Three contigs containing the gene paralogs are represented in light orange with the gene loci represented by the dark-orange small boxes (A-F) in their respective positions (number above the boxes) and orientation, and also with the distance between the genes (numbers between two boxes) in the contigs (A). The corresponding protein encoded by each gene and the variant alleles encoding the proteins Ed-3375891, Ed-3684774, and Ed-444549 used in functional assays are indicated, and the protein similarity compared with Ed-3375891, previously identified as PHPA_79 (B).

The gene *cAvr1-PHPA79-254_1151Kb* possesses two alleles that differ in three nucleotides resulting in two non-synonymous amino acid substitutions (M81I and T174S) in the predicted proteins. The domains predictions in Interproscan for the protein Ed-4444549 identified the signal peptide in the protein encoded by both alleles and one consensus disorder domain in the N-terminal region, between the amino acids 159 and 180, only in the allele containing the serine residue at position 174. However, the Depicter tool predicted the disorder domain in both alleles. The protein Ed-3684774 showed a signal peptide region identified by SignalP, but no other domain was identified. The predicted proteins encoded by the two *cAvr1-PHPA_79* paralog genes of contig 254 were very similar (95% of protein similarity and 97,8% of gene identity in pairwise comparisons). The gene *cAvr1-PHPA79-221_967Kb* was identified on the contig 221 of 71G2h genome, it exhibited three variant alleles, but only two of them correspond to genic variants of the protein Ed-3375891. The sequence analysis of both variant's alleles showed seven polymorphic sites in the gene coding region resulting in synonymous substitution and an additional dinucleotide (TTGGGT to TTAAGT) change observed in the gene sequence of eleven isolates resulting in a non-synonymous amino acid substitution in one of the codons (G125S). A signal peptide is predicted in the protein Ed-3375891 in the N-terminal end, but no other domain was predicted in any of its two variants (Table 2).

Contigs 221 and 254, which harbor three paralogs of the *cAvr1-PHPA_79* gene, were affected by a mutation event that resulted in a large genotypic variation observed in eight large contigs. From 11 isolates virulent on soybean genotypes containing *Rpp1b*, five virulent isolates were homozygotic in all the paralog genes of *cAvr1-PHPA_79*, which means that one allele was observed in its genomes. The other six isolates exhibited two different alleles for each paralog gene, one of these paralog genes with the polymorphism of dinucleotide (an allele of the gene *cAvr1-PHPA79-221_967Kb*) able to encode an alternative allele of the protein Ed-3375891. None of the *P. pachyrhizi* avirulent isolates that were sequenced contained the loss of heterozygosity genotype, but the polymorphism resulting from the dinucleotide change was observed in three of the avirulent isolates (Figure 9, Table 2).

To confirm the presence of dinucleotides polymorphisms in the genic region of *cAvr1-PHPA79-221_967Kb* a fragment of approximately 400 bp including the dinucleotide mutation was amplified by PCR and sequenced in 21 *P. pachyrhizi* isolates (seven virulent and 14 avirulent). The sequences analyzed showed that four virulent isolates were homozygous, corresponding to the isolates with loss of heterozygosity (PHPA 01, 03, 05, and

06) in this genic region, three virulent isolates exhibited the dinucleotide variation in comparison with avirulent isolates that were heterozygotic.

The two other candidates' loci related to virulence exhibited only one polymorphic site in comparison among virulent and avirulent isolates. However, RNA-seq reads mapped on 71G2 genome assembly and gene predictions did not support that these polymorphisms were located in coding regions (Figures 7-8). The candidate *cAvr1-PHPA79-221_977Kb* was located approximately 8,5 Kb apart from the candidate avirulence gene *cAvr1-PHPA79-221_967Kb* (*cAvr1-PHPA_79* from the previous study in PPUFV02) on 71G2h assembly. Virulent genotypes were homozygous A/A and avirulent ones were heterozygous A/G or homozygous G/G (only on isolate PHPA 8) (Figure 8). There was no gene prediction on this locus in any *P. pachyrhizi* annotated genome and the RNA-seq data used in the analyses did not support the expression of genes in this region. For the candidate *cAvr1-60_267Kb* one site polymorphic was also identified in comparison among virulent (A/A) and avirulent (A/C) isolates (Figure 7). This locus encodes the protein ID 492078 of PPUFV02, annotated as a non-secreted protein with a phosphatase 2C-like domain, also present in different species of fungi. For polymorphism analysis, data from RNA-seq transcripts and gene prediction were used and it was demonstrated that the polymorphic site is located in an intron region, and therefore it was not affecting the encoded protein (Figure 7). In analysis, the typical intron-conserved domains recognized by the spliceosome, such as 5' and 3' splicing site, branch point, and polypyrimidine sequences, could be identified in the intron region and the polymorphism was not inside these sequences, suggesting that *cAvr1-PHPA79-221_*967Kb do not correspond to *Avr1*.

### 3.4 Functional analysis of candidate avirulence genes *cAvr5-138_913Kb* and *cAvr1-PHPA_79*

#### 3.4.1 PTI and ETI suppression assays in *N. benthamiana* plants

As some virulent isolates were homozygous for one of the alleles present in avirulent isolates, it was investigated whether *cAvr5-138_913Kb* and the paralogs of *cAvr1-PHPA_79* could induce or suppress the resistance responses in *N. benthamiana* plants. For this purpose, the alleles of the genes *cAvr5-138_913Kb*, *cAvr1-PHPA79-221_967Kb*, *cAvr1-PHPA79-254_1125Kb,* and *cAvr1-PHPA79-254_1151Kb* were transiently expressed using the non-phytopathogenic bacteria *P. fluorescens* Ethan (PfE), and PTI and ETI suppression were analyzed.

For the PTI suppression assay, each candidate *Avr* gene (in pEDV plasmid) was delivered by *P. fluorescens* into *N. benthamiana* leaves, and seven hours later the pathogenic bacterium *P. syringae* pv. *tomato* (Pto) DC3000 was infiltrated. In the assay, the non-pathogenic bacteria activate PTI avoiding the translocation of type III effectors (T3E) by pathogenic bacteria, suppressing the Hypersensitive Response (HR). However, if the Avr candidate in non-pathogenic bacteria is able to suppress PTI response, the ETI response will be activated and the HR response will be observed.

The PTI suppression assays were performed using four gene transformants (*cAvr5-138_913Kb*, *cAvr1-PHPA79-221_967Kb*, *cAvr1-PHPA79-254_1125Kb,* and *cAvr1-PHPA79-254_1151Kb*). The experiments were performed twice and in both PTI suppression, evaluated by the presence of HR in the co-infiltrated area, was lower than 40%, demonstrating that the candidate avirulence genes were not PTI suppressors in *N. benthamiana.* The effector PTI suppressor EC23 from *P. pachyrhizi* was used as a positive control, and it was able to suppress the PTI response in more than 50% of the infiltrated leaves.

ETI suppression was measured by the capacity of the candidate effector protein to prevent the hypersensitivity response triggered by the AvrB recognition on the *N. benthamiana* after its co-inoculation. However, none of our candidate effector proteins suppressed the HR response in more than 50% of the inoculated leaves, not being considered as ETI suppressors. The best results, HR suppression in 40% of the evaluated leaves, were obtained for the protein encoded by gene *cAvr1-PHPA79-254_1151Kb* (Ed-4444549).

### 3.4.2 Co-expression of candidate Avr5 and corresponding resistance gene in *N. benthamiana* plants

Recently, the soybean resistance gene *Rpp5* and a second gene, denominated *7900*, encoding a protein that possibly interacts with the Rpp5 to trigger immunity to *P. pachyrhizi*, were mapped in soybean genotypes and cloned (unpublished data). To verify the direct recognition of the candidate Avr5 proteins by Rpp5 and the association with 7900, *N. benthamiana* leaves were co-inoculated with *A. tumefaciens* transformed with genes in the following combinations: *Rpp5* and *cAvr5-138_913Kb,* and *Rpp5*, *cAvr5-138_913Kb* and *7900*. It was expected that Avr-NLR recognition would result in HR or leaf chlorosis. The inoculation of the transformant clones alone or in combinations did not result in any macroscopic reaction. Leaves inoculated with *A. tumefaciens* transformed with the *BAR* gene,

which confers resistance to the herbicide glufosinate was used as control of the transient transformation and resulted in herbicide resistance when the herbicide was applied on the infiltrated area of the leaves. These results indicated that the candidate protein encoded by the *cAvr5-138_913Kb* gene was not recognized by the Rpp5 when expressed in *N. benthamiana*.

### 3.4.3 Subcellular localization of Avr1 candidates proteins in *N. benthamiana* plants

*N. benthamiana* leaves were transiently transformed for expression of the proteins encoded by *cAvr1-PHPA79* alleles (Ed-3375891, Ed-3684774, and Ed-4444549) without their predicted signal peptide and fused with the GFP. The Ed-3375891 protein isoforms (Al1: present in all isolates; Al2: present only in isolates heterozygotic for the dinucleotide variation), encoded by gene *cAvr1-PHPA79-221_967Kb*, and proteins Ed-3684774 and Ed-4444549 isoforms (Al1: present in all isolates; Al2: present only in isolates heterozygotic isolates), encoded by genes *cAvr1-PHPA79-254_1125Kb* and *cAvr1-PHPA79-254_1151Kb*, displayed co-localization with the nuclei and possibly with the cell membrane, not co-localized with the vacuole marker (Figures 11-13).

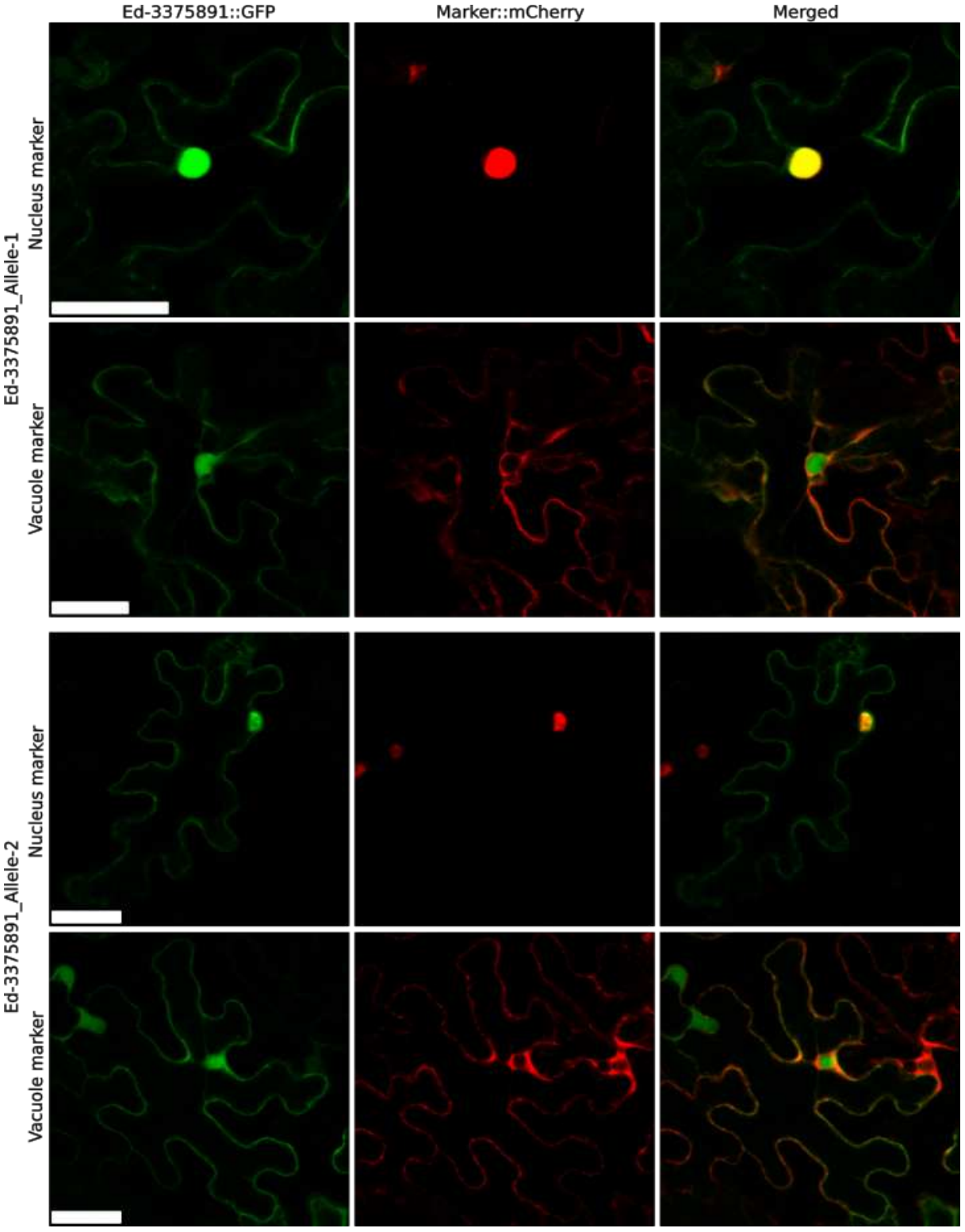



**Figure 11:** Subcellular localization of isoforms 1 and 2 of candidate avirulence protein Ed-3375891 in *N. benthamiana*. The green color indicates the localization of the candidate Avr protein N-terminal fused with GFP, the red color indicates the localization of the nuclear marker (pMP90::*WWP1-mCherry* - Calil et al., 2018) or the vacuolar marker (pMP90::*Atγ-TIP-mCherry* - Saito et al., 2002), and the yellow color indicates co-localization between the avirulence protein candidate and the marker protein. The constructions were co-infiltrated and 72 hours after inoculations were visualized by confocal microscopy. GFP and mCherry were excited at 488 and 561 nm, respectively. GFP (green) and mCherry (red) fluorescence were collected at 505 to 525 nm and 580 to 620 nm, respectively. White bar = 50 µm.

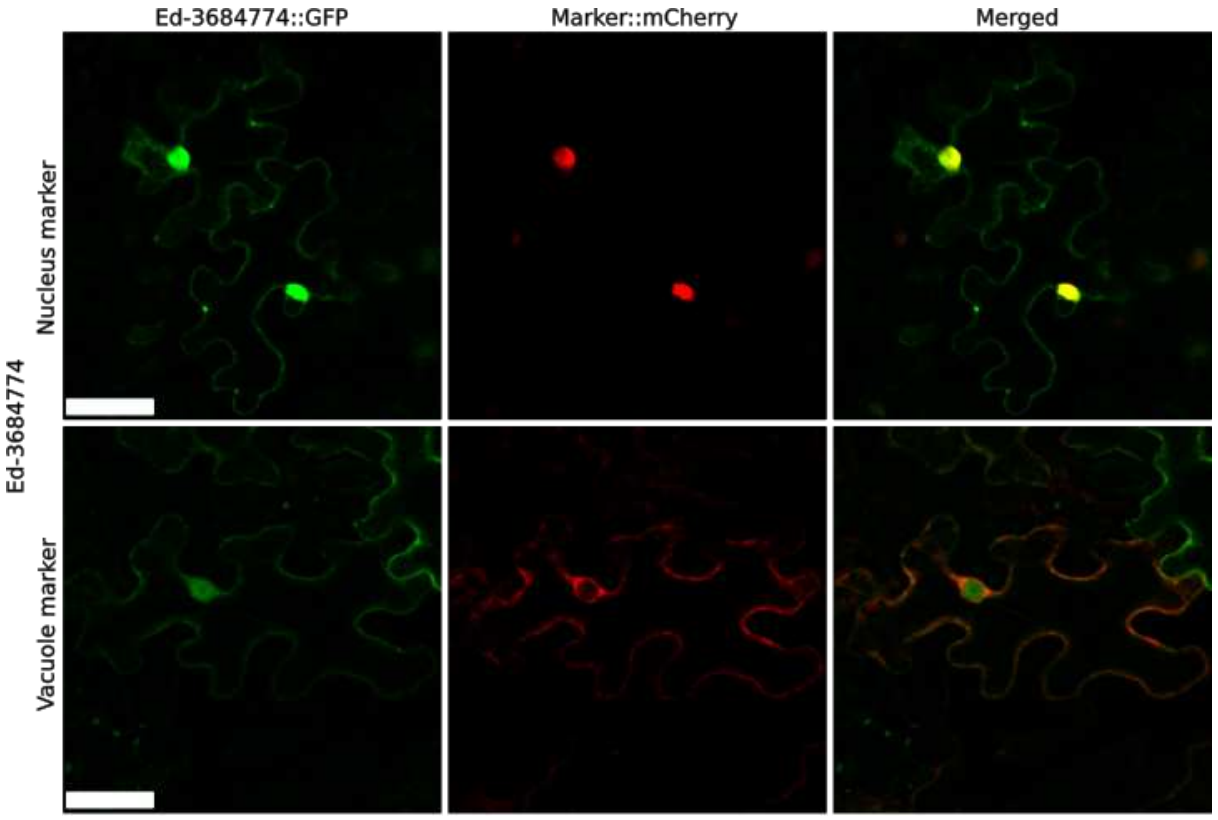



**Figure 12:** Subcellular localization of the candidate avirulence protein Ed-3684774 in *N. benthamiana* leaves. The green color indicates the localization of the candidate Avr protein N-terminal fused with GFP, the red color indicates the localization of the nuclear marker (pMP90::*WWP1-mCherry* - Calil et al., 2018) or the vacuolar marker (pMP90::*Aty-TIP-mCherry* - Saito et al., 2002), and the yellow color indicates co-localization between the avirulence protein candidate and the marker protein. The constructions were co-infiltrated and 72 hours after inoculations were visualized by confocal microscopy. GFP and mCherry were excited at 488 and 561 nm, respectively. GFP (green) and mCherry (red) fluorescence were collected at 505 to 525 nm and 580 to 620 nm, respectively. White bar = 50 µm.

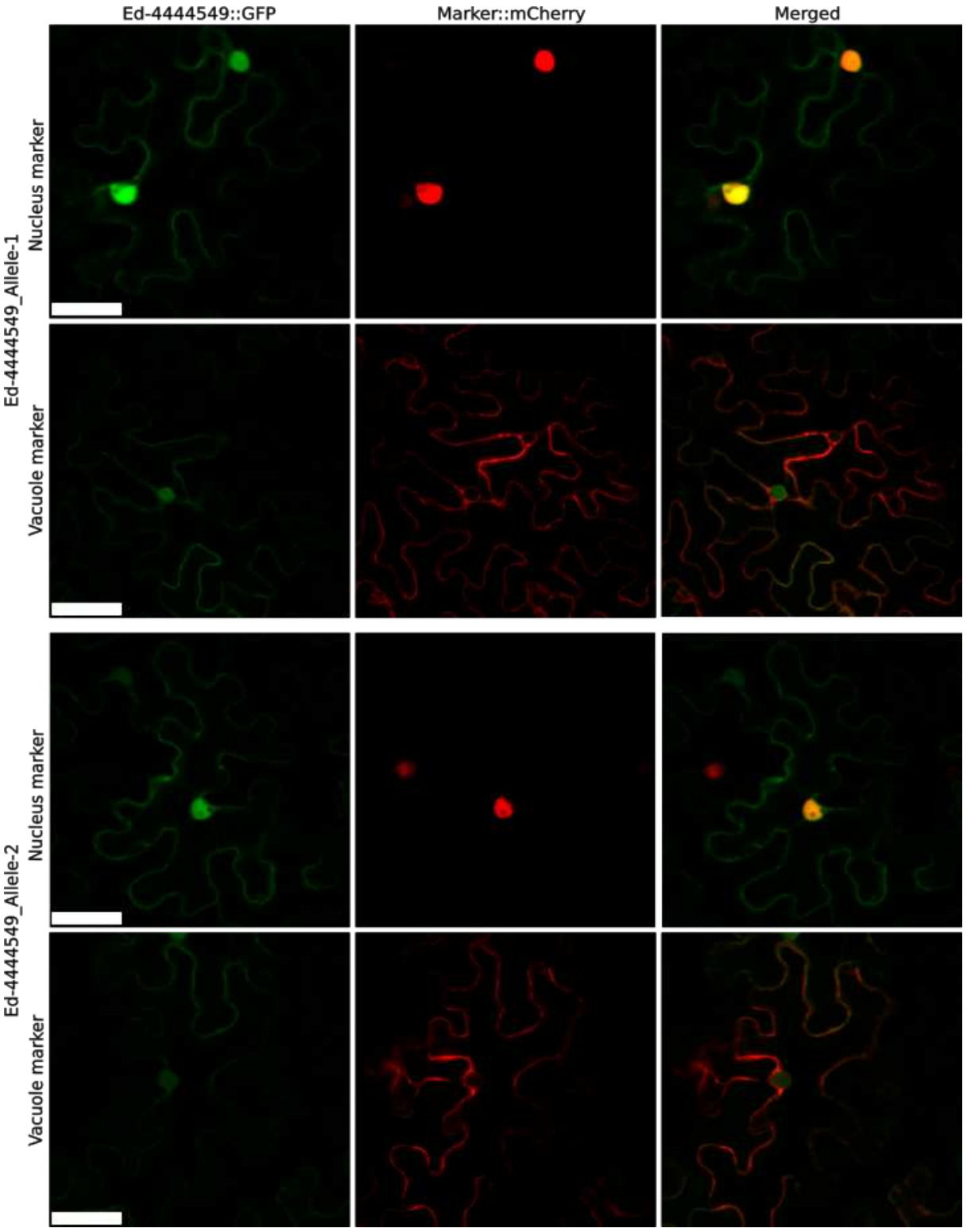



**Figure 13:** Subcellular localization of isoforms 1 and 2 of candidate avirulence protein Ed-4444549 in *N. benthamiana*. The green color indicates the localization of the candidate Avr protein N-terminal fused with GFP, the red color indicates the localization of the nuclear marker (pMP90::*WWP1-mCherry* - Calil et al., 2018) or the vacuolar marker (pMP90::*Atγ-TIP-mCherry* - Saito et al., 2002), and the yellow color indicates co-localization between the avirulence protein candidate and the marker protein. The constructions were co-infiltrated and 72 hours after inoculations were visualized by confocal microscopy. GFP and mCherry were excited at 488 and 561 nm, respectively. GFP (green) and mCherry (red) fluorescence were collected at 505 to 525 nm and 580 to 620 nm, respectively. White bar = 50 µm.

## 4. DISCUSSION

The recent advance in DNA sequencing technologies and assembly tools especially in the last decade allowed the assembly of whole genomes from rust species with complex genomes such as *Hemileia vastatrix* (Porto et al., 2019)*, Austropuccinia psidii* (Tobias et al., 2021) and *P. pachyrhizi* (Gupta et al., 2023). These genomes encode thousands of potential effector genes that typically share minimal or no sequence homology to other proteins with known functions. Some of these effectors can be recognized by resistance (R) proteins and are therefore also called avirulence (Avr) proteins. Avr recognition triggers a suite of defense-related host responses to fight off pathogen infection including the hypersensitive response (HR) that leads to cell death at the site of infection and halts pathogen growth (Jones & Dangl, 2006; Dangl et al., 2013). However, pathogens continuously evolve to escape recognition via mutation of *Avr* genes. Therefore, the identification and characterization of genes coding Avr proteins and the understanding of Avr gene diversity in pathogen populations is critical for the effective deployment of R genes in breeding or gene stacking. In the case of *Phakopsora pachyrhizi*, its genome complexity, the dikaryotic nature of its asexual stage, the absence of functional sexual reproduction, and the absence of genetic transformation systems, associated with its obligate parasitism impose a severe limitation to the identification of *Avr* genes using genetic mapping, as well as gene cloning and functional analysis.

Trying to overcome these technical hurdles, the most advances in sequencing technologies and bioinformatics were combined with a genome-wide approach to identify candidate Avr among the thousands of candidate effector genes annotated on the *P. pachyrhizi* genome. First, it was selected and purified a “near-isogenic” virulent and avirulent natural mutants. Second, it was generated a reference diploid assembly for the *P. pachyrhizi* isolate 71G2 genome, from which a *Rpp1b* virulent variant was initially identified. During this research, other *P. pachyrhizi* genomes became available. However, the 71G2 assembled genome was used as the reference genome for our association studies, to minimize genomic differences not related to the gain of virulence, that would arise from other random mutations. Considering that the avirulence genes are generally dominant (Dodds et al., 2020) and it was possible to purify natural virulent isolates from avirulent isolates during the isolates multiplication at the laboratory, it was hypothesized that the avirulent isolates should be heterozygous at *Avr1* and *Avr5* loci. Mutation of the avirulence allele or loss of heterozygosity could generate virulent isolates from an avirulent. To identify genetic changes that could be

associated with this gain of virulence, initially pairwise comparisons of short-read sequences from two *Rpp5* avirulent isolates against one virulent isolate and two *Rpp1b* avirulent isolates aligned against the reference haplotype 71G2h were done, looking for homozygous polymorphism in the virulent isolates and corresponding heterozygous polymorphism in the avirulent isolates. A huge number of polymorphisms with these features was initially identified, mainly in regions rich in retrotransposable elements. Due to the different ages and similarities of these elements (Gupta et al., 2023), it was observed cross-mapping of short read sequences from different elements, generating pseudo-heterozygosity and/or polymorphisms. To narrow down the analysis to regions containing effector genes, the initially identified polymorphisms that mapped to candidate effector genes were filtered. For this purpose, a manually curated database with expressed genes of *P. pachyrhizi* predicted to encode secreted proteins was created. Using this approach, four candidates for *Avr5* were identified, but it was not possible to identify candidate genes for *Avr1*. Probably, more than one polymorphism present in different positions in different isolates is associated with the gain of virulence on *Rpp1b,* all of them are affecting the recognition of the encoded protein, as reported for different avirulence genes in different pathosystems (Petit-Houdenot & Fudal, 2017). Therefore, a manual screening was done to identify genes with variations among *Rpp1b* virulent and avirulent isolates of *P. pachyrhizi*. In this process, a large genomic region with loss of heterozygosity, affecting eight large contigs, was identified in five *Rpp1b* virulent isolates. Considering the large region affected impacting diverse genes and the strong impact of the complete loss of one allele of a gene, a manual screening for candidates *Avr1* was performed in these contigs. With this new approach, three variant loci possibly related to the virulence characteristic were identified and analyzed in detail. Limiting the analyzed genes to that predicted to encode a secreted protein could reduce the number of genes to be analyzed and the time to find a candidate gene, but it also could fail if the database does not contain the correct avirulence gene, a situation that could happen if, for example, the gene was not predicted in the genome annotation or if the encoded protein was not identified as secreted due to signal peptide prediction error or if the protein is secreted by an unconventional secretion pathway (in this case there will be no signal peptide) (Sonah et al., 2016; Rabouille, 2017).

Variation sites from each candidate avirulence gene were manually checked to confirm the polymorphism among the 37 *P. pachyrhizi* DBN-seq sequenced and aligned genomes, including eleven *Rpp1b* virulent isolates and the two unique *Rpp5* virulent isolates. In this analysis, different kinds of variation affecting the candidate avirulence genes were observed

and they were organized into four classes for the variation point: 1) bi-allelic genes common to all isolates. One of the two alleles is present among all isolates, and the second allele has one nucleotide variant between virulent and avirulent isolates (*cAvr5-47_1128Kb*, *cAvr5-134_549Kb,* and *cAvr5-310_66Kb*); 2) One SNP among virulent and avirulent isolates, where the virulent isolates are homozygotic and the avirulent ones are heterozygotic (*cAvr1-60_267Kb* and *cAvr1-221_9777Kb*); 3) Loss of heterozygosity in the virulent isolates (*cAvr5-138_913Kb*); 4) Association of mutation events in paralog genes: loss of heterozygosity in one paralog gene for virulent isolates, and a dinucleotide variation affecting virulent isolates that were not affected by the loss of heterozygosity, but also affecting four avirulent isolates (*cAvr1-PHPA79-221_967Kb* combined with *cAvr1-PHPA79-254_1151Kb*).

Based on these four classes (1-4), some possibilities could explain how the genotypic variation could affect the resistance phenotype in cases of avirulence gene changes to avoid resistance genes recognition (Chen et al., 2017; Petit-Houdenot & Fudal, 2017; Ortiz et al., 2022). In the first class, considering these bi-allelic genes as possible avirulence alleles *Avr/avr*, a single amino acid change in the Avr peptide of the virulent isolates could affect its recognition by the resistance protein, avoiding the resistance phenotype. The second class is similar to the first, but the variation depends on the variant allele present only in the heterozygotic avirulent isolates (*Avr*/*avr*) where the allele encoding the Avr factor triggers resistance. In both cases, a single amino acid variation in the encoded protein Avr or avr would determine the resistance or susceptibility, as reported for several avirulence genes such as the *AvrSr50* and the *AvrL567* of *Puccinia graminis* f. sp. *tritici* and *Melampsora lini*, respectively (Wang et al., 2007; Ortiz et al., 2022). The third class corresponds to a loss of heterozygosity which generally is observed in asexual or clonally propagated species. It would be the consequence of a mutation event when one genomic region including one of the alleles of a previous heterozygotic individual is lost and the individual turns homozygotic (Smukowski Heil, 2023). If the region encoding the Avr factor is lost in this event, it could explain the gain of virulence of the isolate, as recently reported for the *AvrSr50* of *Puccinia graminis* f. sp. *tritici* (Chen et al., 2017). The fourth possibility involves the combined analysis of the *PHPA_79* paralogs, evolving the loss of heterozygosity for one allele of the gene *cAvr1-PHPA79-254_1151Kb* and the dinucleotide mutation in the gene *cAvr1-PHPA79-221_967Kb* causing some gain of virulence. After the gene duplication that generated paralogs, both genes exhibited the same function, and with time, the action of evolutive mechanisms on gene sequences could affect their function and expression. With the presence of copies, gene sub-functionalization (a division of the initial function) or new

function development (neofunctionalization) usually occurs (Li et al., 2005; Wapinski et al., 2007). In some cases, mostly in sub-functionalization, paralogs could interact with each other to maintain the initial gene function (Lynch & Force, 2000; Wapinski et al., 2007). Furthermore, the occurrence of avirulence genes organized in genic families, exhibiting variable number of copies and alleles, have been reported for the avirulence genes such as the genes *AvrL567*, *AvrM*, and *AvrL2* of *Melamposara lini* (Wang et al., 2007; Ve et al., 2013; Wu et al., 2019), which may also occur for *P. pachyrhizi Avr* genes. Considering the possibility of interaction between paralogs, the combination of polymorphisms in the paralogs of *PHPA_79* could be determining the virulence. However, this situation is very improbable principally due to the presence of the dinucleotide mutation in four avirulent isolates, drastically reducing the probability of this gene be the *Avr1*. No other genotypic variations, that could explain their genotypic deflection, were identified in these four avirulent isolates, hence deepen analysis has to be done in future studies.

Additional bioinformatic analyses were performed to determine the impact of the genomic variation in the candidate avirulence genes and in their encoded proteins to verify if the proteins have characteristics usually described for an effector protein. Among the factors evaluated were to be secreted, species or genus-specific, have a small size (less than 300 amino acids), and exhibit a high number of cysteine residues (Stergiopoulos & de Wit, 2009; Lorrain et al., 2019).

At the beginning of the analysis, two *Avr1* of the candidates (*cAvr1-60_267Kb* and *cAvr1-221_977Kb*) did not exhibit polymorphism in coding regions. The *cAvr1-60_267Kb* candidate was predicted to encode a non-secreted Phosphatase protein containing the 2C-like domain, a protein present in diverse organisms, including eukaryotes and prokaryotes. Its function is to remove phosphate groups from diverse proteins, acting as a key regulator component in signaling pathways. However, the protein prediction and the RNA-seq reads alignment support that the polymorphic site is located in an intronic region, but they were not affecting the splicing sites recognized by spliceosome during the processing.

The locus corresponding to *cAvr1-221_977Kb* showed a polymorphic site that perfectly separates in virulent and avirulent isolates, but it was not possible to identify the gene or RNA-seq transcripts associated with this variation. The RNA-seq data coverage could be affected by the gene expression, which in non-constitutive genes could be induced or repressed in specific conditions and not represent all expressed genes (Marguerat and Bahler., 2009). RNA-seq data obtained from the PPUFV02 files available at the Mycocosm combining transcripts from different moments of the interaction of *P. pachyrhizi*-soybean were used in

this analysis, but maybe these conditions were not sufficient to capture the correct moment of expression of some genes. The gene prediction tools also could fail to predict some genes due to the low similarity of specie-specific effector genes and sometimes the absence of expression signals not captured in the RNA-seq, features that affect the success of most of the prediction tools that use annotated proteins from other species and external pieces of information (such as RNA-seq data) to predict genes (Brůna et al., 2020).

The third candidate *Avr1* was previously identified in an EST library (unpublished data) as *PHPA_79*, but before this analysis, it was not known about the presence of three closely related paralogs. Interestingly, this candidate gene exhibited genetic variations when virulent/avirulent isolates were compared. The three paralog genes showed loss of heterozygosity, they are *P. pachyrhizi* specific, contain multiple paralog genes, and encode a small secreted protein (180 amino acids) with a high number of cysteine residues. Based on all these interesting characteristics the *PHPA_79* was considered the best candidate *Avr1* to proceed with functional analysis.

The candidates for *Avr5* were also analyzed and variations in genic regions of the *P. pachyrhizi* isolates were observed. The analysis of two loci candidates: *cAvr5-47_1128Kb* and *cAvr5-134_549Kb* indicated genes with unknown functions encoding proteins with transmembrane and membrane-bound domains, respectively, both characteristics not expected in Avr proteins. The *cAvr5-47_1128Kb* gene has homologs restricted to rust fungi, encoding a hypothetical protein, and *cAvr5-134_549Kb* is exclusive to *P. pachyrhizi*, not being identified in other species. The taxonomic specificity and unknown function features are usually used in the screening of effector genes due to their specific interaction with the host and possible fast evolution, but the presence of transmembrane or membrane-bound domains often are used to discard possible candidates (Stergiopoulos & de Wit, 2009; Sperschneider et al., 2017; Lorrain et al., 2019), which reduced the chance of these two candidates being the Avr5 protein. Despite that, both candidates (*cAvr5-47_1128Kb* and *cAvr5-134_549Kb*) were not discarded, since they exhibited genotype/phenotype association, specificity to rust, and also membrane association predictions may not be correct.

The third candidate *cAvr5-310_66Kb* encodes an alpha/beta hydrolase protein containing a serine carboxypeptidase catalytic domain, with serine carboxypeptidase activity, also annotated in other basidiomycetes fungi. Serine peptidases were reported as virulence proteins for the fungi *Fusarium eumartii* (Olivieri et al., 2002) and *F. oxysporum* f. sp. *lycopersici* (Jashni et al., 2015) and for some plant parasitic nematodes such as *Aphelenchoides besseyi* (Huang et al., 2022) and *Radopholus similis* (Huang et al., 2017). In

both studies with *Fusarium* spp., the serine peptidases degrade pathogenesis-related proteins secreted by plants, as chitinases, and support the fungi infection. In nematode studies, the function of the serine carboxypeptidases is not clear, but in both cases, RNAi-treated nematodes with serine peptidases were less pathogenic. Although *cAvr5-310_66Kb* is not a protein exclusive to *P. pachyrhizi*, the possible action as a plant pathogen effector and the polymorphism among virulent/avirulent isolates, suggested that this gene candidate should be used to advance the studies of avirulence.

The fourth candidate *Avr5* identified in our comparative analysis possibly encodes a peptide with 33 amino acids, that was bi-allelic in avirulent isolates and monoallelic in avirulent ones. The possible ORF was recovered from RNA transcripts supporting the expression of the gene and delimiting the transcript sequence. The peptides possibly encoded by the alleles of *cAvr5-138_913Kb* contain four polymorphic amino acids and did not show similarity with other proteins, no predicted domains, and no signal peptide for secretion. The absence of signal peptide indicates that this peptide is not secreted by the classical secretion route involving the endoplasmic reticulum and Golgi apparatus. However diverse secreted proteins, including plant pathogens effectors, were described as being secreted by different pathways denominated unconventional protein secretion - UPS (Giraldo et al., 2013; Liu et al., 2014; Rabouille, 2017; Krombach et al., 2018). The small size (33 aa) of the peptide could facilitate its secretion for an alternative pathway and to cause a significant impact on the interaction, as reported for the *Cladosporium fulvum* Avr2 (Luderer et al., 2002) and Avr9 (van den Hooven, et al., 2001), *Colletotrichum gloeosporioides* CgDN3 with 58, 28 and 54 amino acids, respectively, in the mature protein (Stephenson et al., 2000). The significant difference between the two alleles of the gene *cAvr5-138_913Kb*, the impact of the complete loss of one allele in virulent isolates, and a species-specific sequence highlight this candidate as the possible *Avr5*.

The analysis for the fourth *Avr5* candidate and the paralog genes to *PHPA_79* candidate *Avr1* suggested that these would be the most likely candidates for *Avr*s genes. Therefore, the candidate *Avr1* (paralogs of *PHPA_79*) and one *Avr5* (*cAvr5-138_913Kb*) were selected for additional analysis. Functional analyses were performed by heterologous expression in *N. benthamiana* using *A. tumefaciens* and *P. fluorescens* EtHAn for transient transformation or expression, respectively. The determination of the cellular targets of pathogens effectors in model plants could be used to infer where the pathogen protein can act in its natural host and to predict the effector function (Sperschneider *et al.*, 2017). So, confocal microscopy was used to determine the subcellular localization of the proteins

encoded by the paralog genes of candidate *cAvr1-PHPA_79*. The proteins Ed-3684774, Ed-4444549, and Ed-3375891, encoded by the *cAvr1-PHPA_79* paralog genes, had the same subcellular localization, in the nucleus and possibly the plasmatic membrane. Nuclear effectors are expected to move initially along cytoplasm and enter the nucleus to suppress the host defense response or reprogram the transcription of the host defense genes, while nucleo-cytoplasmic effectors could execute the cited function when in the nucleus or act on cytosolic proteins, disrupting the defense surveillance or affecting the signal transduction on kinase cascades and consequently altering the plant metabolism, supporting the infection and propagation of the pathogen (Prasad et al., 2019; Figueroa et al., 2021). The localization in more than one subcellular compartment also could be associated with different periods of the infection or be an effect of the expression in a heterologous system.

The *Rpp5* resistance gene and its auxiliary gene *7900* were recently cloned (unpublished data) and associated with the effective resistance on soybeans. The Agrobacterium-mediated transformation was used for co-expression of the resistance gene and the candidate cognate avirulence gene *cAvr5-138_913Kb*, with and without the gene *7900.* In both situations, the co-expression did not result in the induction of HR, indicating that the *cAvr5-138_913Kb* was not recognized by the Rpp5 or that this combination was not sufficient to trigger the resistance response in *N. benthamiana.* The system does not indicate protein interaction, maybe because it does not contain all the factors necessary for the protein-protein interaction, such as other pathogen effectors or other components of the plant immune system.

In the heterologous expression assay, using *P. fluorescens* EtHAn the suppression of the PTI or ETI was not observed by candidate effector genes. However, proteins without suppression ability (PTI or ETI) can be involved in other important processes to host manipulation and pathogen infection such as obtaining nutrients, development of the pathogen, or acting in a manner dependent on other host/pathogen proteins and additional effectors (Rovenich et al., 2014).

Unfortunately, the effector activity or the recognition of the candidate effectors using heterologous expression in *N. benthamiana* were not confirmed. These initial results do not indicate that our candidates *PHPA_79* and *cAvr5-138_913Kb* are the genes *Avr1* and *Avr5*, respectively. However, the heterologous expression results are not sufficient to discard the candidates because some processes or interactions only occur in specific biological settings with the pathogen and its host (Lorrain et al., 2018). In general, there are several limitations to performing functional analysis in the pathosystem *P. pachyrhizi*-soybean because there is no

established protocol for the fungus transformation and cultures without the host, and the soybean plant also is not easily infiltrated and transformed to perform transiently expression in its leaves. For these reasons, heterologous assays using bacteria for transient expression of proteins in *N. benthamiana* are broadly used. However, the differences from the natural system are significant and these assays usually are used just for an initial screening with limited conclusions because heterologous systems may not contain factors necessary for the interaction process as host or pathogen-specific proteins, the adequate expression of the genes or protein modifications in the correct time and circumstances that could happen in the natural system (Lorrain et al., 2018). For these reasons, in further studies, functional assays expressing the candidate effector proteins in soybean protoplasts of resistant and susceptible genotypes will be performed to reduce possible limitations of the heterologous assay related to the non-host plant.

The genomic analyses were important for the identification of candidate avirulence and effector genes in *P. pachyrhizi* with comparative analysis. However, future studies must be done to improve the detection of candidate avirulence genes and the correct ORF of them because we have signs that some genes, usually small sequences or species-specific genes, were not predicted (such as the candidate *cAvr5-138_913Kb*) or were incorrectly predicted (such as the candidate genes *PHPA_79*) even when the genome annotation of three different isolates was used, and for this reason, some candidate effector genes could be missing. As the gene prediction could be incomplete, it will need to improve the analysis of variants using variant call tools with multiple genotypes to not be limited to a database that could miss some genes. Signs of alternative splicing in some multi-exon genes with partial degenerated splicing sites were also identified, suggesting variable coverage of RNA-seq reads for different exons of the same gene. In some cases, alternative splicing is induced on stress conditions during the fungi-host interaction (Sieber et al., 2018) and for this reason, some effector genes could be affected, and if it does happen, splicing variants and genotypic variations in introns could be associated with phenotypic variations. Alternative splicing enhances the complexity of the transcriptome, generating alternative transcripts and proteins from a single multi-exon gene (Matlin et al., 2005) and also could regulate gene expression (Ge & Porse, 2014). This process is reported in diverse species of fungi (Kupfer et al., 2004; Sieber et al., 2018; Muzafar et al., 2021), but it was not explored in rust fungi. For this reason, the transcriptomic analysis also has to be implemented in future analyses.

## 5. CONCLUSIONS

In conclusion, the comparative analysis resulted in the identification of two candidates *Avr1*, four candidates *Avr5*. The analysis of variant calls had more success than the use of predicted secreted genes, indicating that the avirulence gene could be secreted by an unconventional secretory pathway. Heterologous expression assays of two candidate avirulence factors did not indicate their function as plant immunity suppressors and the expression of the *Avr5* with *Rpp5* did not trigger HR in *N. benthamiana*. These results could be associated with technique limitations.

## 6. REFERENCES


Almagro Armenteros, J. J., Tsirigos, K. D., Sønderby, C. K., Petersen, T. N., Winther, O., Brunak, S., ... & Nielsen, H. (2019). SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nature Biotechnology*, 37 (4), 420-423.

Badel, J. L., Piquerez, S. J., Greenshields, D., Rallapalli, G., Fabro, G., Ishaque, N., & Jones, J. D. (2013). In planta effector competition assays detect *Hyaloperonospora arabidopsidis* effectors that contribute to virulence and localize to different plant subcellular compartments. *Molecular plant-microbe interactions*, 26 (7), 745-757.

Badet, T., & Croll, D. (2020). The rise and fall of genes: origins and functions of plant pathogen pangenomes. *Current opinion in plant biology*, 56, 65-73.

Baltrus, D. A. (2017). Adaptation, specialization, and coevolution within phytobiomes. *Current opinion in plant biology*, 38, 109-116.

Barik, A., Katuwawala, A., Hanson, J., Paliwal, K., Zhou, Y., & Kurgan, L. (2020). DEPICTER: intrinsic disorder and disorder function prediction server. *Journal of molecular biology*, 432 (11), 3379-3387.

Bonde, M. R., Nester, S. E., Austin, C. N., Stone, C. L., Frederick, R. D., Hartman, G. L., & Miles, M. R. (2006). Evaluation of virulence of *Phakopsora pachyrhizi* and *P. meibomiae* isolates. *Plant Disease*, 90 (6), 708-716.

Bonde, M. R., Nester, S. E., Berner, D. K., Frederick, R. D., Moore, W. F., & Little, S. (2008). Comparative susceptibilities of legume species to infection by *Phakopsora pachyrhizi*. *Plant Disease*, 92 (1), 30-36.

Bonifacino, J. S., & Glick, B. S. (2004). The mechanisms of vesicle budding and fusion. *Cell*, 116 (2), 153-166.

Bromfield, K. R., & Hartwig, E. E. (1980). Resistance to soybean rust and mode of inheritance. *Crop Science*, 20 (2), 254-255.

Brooks, A. W., Kohl, K. D., Brucker, R. M., van Opstal, E. J., & Bordenstein, S. R. (2016). Phylosymbiosis: relationships and functional effects of microbial communities across host evolutionary history. *PLoS biology*, 14 (11), e2000225.

Calil, I. P., Quadros, I. P., Araújo, T. C., Duarte, C. E., Gouveia-Mageste, B. C., Silva, J. C. F., ... & Fontes, E. P. (2018). A WW domain-containing protein forms immune nuclear bodies against begomoviruses. *Molecular plant*, 11 (12), 1449-1465.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC bioinformatics*, 10, 1-9.

Chen, J., Upadhyaya, N. M., Ortiz, D., Sperschneider, J., Li, F., Bouton, C., ... & Dodds, P. N. (2017). Loss of *AvrSr50* by somatic exchange in stem rust leads to virulence for *Sr50* resistance in wheat. *Science*, 358 (6370), 1607-1610.

Childs, S. P., King, Z. R., Walker, D. R., Harris, D. K., Pedley, K. F., Buck, J. W., ... & Li, Z. (2018). Discovery of a seventh *Rpp* soybean rust resistance locus in soybean accession PI 605823. *Theoretical and Applied Genetics*, 131, 27-41.

Cooper, B., Campbell, K. B., Beard, H. S., Garrett, W. M., & Islam, N. (2016). Putative rust fungal effector proteins in infected bean and soybean leaves. *Phytopathology*, 106 (5), 491-499.

Cuomo, C. A., Bakkeren, G., Khalil, H. B., Panwar, V., Joly, D., Linning, R., ... & Fellers, J. P. (2017). Comparative analysis highlights variable genome content of wheat rusts and divergence of the mating loci. *G3: Genes, Genomes, Genetics*, 7 (2), 361-376.

Dangl, J. L., Horvath, D. M., & Staskawicz, B. J. (2013). Pivoting the plant immune system from dissection to deployment. *Science*, 341 (6147), 746-751.

de Carvalho, M. C. D. C., Costa Nascimento, L., Darben, L. M., Polizel‐Podanosqui, A. M., Lopes‐Caitar, V. S., Qi, M., ... & Marcelino‐Guimarães, F. C. (2017). Prediction of the in planta *Phakopsora pachyrhizi* secretome and potential effector families. *Molecular plant pathology*, 18 (3), 363-377.

De Guillen, K., Lorrain, C., Tsan, P., Barthe, P., Petre, B., Saveleva, N., ... & Hecker, A. (2019). Structural genomics applied to the rust fungus *Melampsora larici-populina* reveals two candidate effector proteins adopting cystine knot and NTF2-like protein folds. *Scientific Reports*, 9 (1), 1-12.

Di Tommaso, P., Moretti, S., Xenarios, I., Orobitg, M., Montanyola, A., Chang, J. M., ... & Notredame, C. (2011). T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucleic acids research*, 39, W13-W17.

Dodds, P. N., Lawrence, G. J., Pryor, A., & Ellis, J. G. (2020). Genetic analysis and evolution of plant disease resistance genes. *Molecular Plant Pathology*, 88-107.

Fabro, G., Steinbrenner, J., Coates, M., Ishaque, N., Baxter, L., Studholme, D. J., ... & Jones, J. D. (2011). Multiple candidate effectors from the oomycete pathogen *Hyaloperonospora arabidopsidis* suppress host plant immunity. *PLoS pathogens,* 7 (11), e1002348.

Figueroa, M., Upadhyaya, N. M., Sperschneider, J., Park, R. F., Szabo, L. J., Steffenson, B., ... & Dodds, P. N. (2016). Changing the game: using integrative genomics

to probe virulence mechanisms of the stem rust pathogen *Puccinia graminis* f. sp. *tritici*. *Frontiers in Plant Science*, 7, 205.

Figueroa, M., Dodds, P. N., & Henningsen, E. C. (2020). Evolution of virulence in rust fungi—multiple solutions to one problem. *Current opinion in plant biology*, 56, 20-27.

Figueroa, M., Ortiz, D., & Henningsen, E. C. (2021). Tactics of host manipulation by intracellular effectors from plant pathogenic fungi. *Current Opinion in Plant Biology*, 62, 102054.

Garcia, A., Calvo, É. S., de Souza Kiihl, R. A., Harada, A., Hiromoto, D. M., & Vieira, L. G. E. (2008). Molecular mapping of soybean rust (*Phakopsora pachyrhizi*) resistance genes: discovery of a novel locus and alleles. *Theoretical and Applied Genetics*, 117 (4), 545-553.

Ge, Y., & Porse, B. T. (2014). The functional consequences of intron retention: alternative splicing coupled to NMD as a regulator of gene expression. *Bioessays*, 36(3), 236-243.

Giraldo, M. C., Dagdas, Y. F., Gupta, Y. K., Mentlak, T. A., Yi, M., Martinez-Rocha, A. L., ... & Valent, B. (2013). Two distinct secretion systems facilitate tissue invasion by the rice blast fungus *Magnaporthe oryzae*. *Nature Communications*, 4 (1), 1-12.

Goellner, K., Loehrer, M., Langenbach, C., Conrath, U. W. E., Koch, E., & Schaffrath, U. (2010). *Phakopsora pachyrhizi*, the causal agent of Asian soybean rust. *Molecular plant pathology*, 11 (2), 169-177.

Gupta, Y. K., Marcelino-Guimarães, F. C., Lorrain, C., Farmer, A. D., Haridas, S., Ferreira, E. G. C., ... & van Esse, H. P. (2023). Major proliferation of transposable elements shaped the genome of the soybean rust pathogen *Phakopsora pachyrhizi*. *Nature Communications*, 14, 1835.

Höfgen, R., & Willmitzer, L. (1988). Storage of competent cells for *Agrobacterium* transformation. *Nucleic acids research*, 16 (20), 9877.

Huang, X., Xu, C. L., Chen, W. Z., Chen, C., & Xie, H. (2017). Cloning and characterization of the first serine carboxypeptidase from a plant parasitic nematode, *Radopholus similis*. *Scientific reports*, 7 (1), 4815.

Huang, X., Chi, Y., Birhan, A. A., Wei, Z, Qi, R., & Peng, D. (2022). The new effector *AbSCP1* of foliar nematode (*Aphelenchoides besseyi*) is required for parasitism rice. *Journal of Integrative Agriculture*, 21(4), 1084-1093.

Hyten, D. L., Hartman, G. L., Nelson, R. L., Frederick, R. D., Concibido, V. C., Narvel, J. M., & Cregan, P. B. (2007). Map location of the *Rpp1* locus that confers resistance to soybean rust in soybean. *Crop Science*, 47 (2), 837-838.

Hyten, D. L., Smith, J. R., Frederick, R. D., Tucker, M. L., Song, Q., & Cregan, P. B. (2009). Bulked segregant analysis using the GoldenGate assay to locate the *Rpp3* locus that confers resistance to soybean rust in soybean. *Crop Science*, 49 (1), 265-271.

Innes, R. W., Bisgrove, S. R., Smith, N. M., Bent, A. F., Staskawicz, B. J., & Liu, Y. C. (1993). Identification of a disease resistance locus in *Arabidopsis* that is functionally homologous to the *RPG1* locus of soybean. *The Plant Journal*, 4 (5), 813-820.

Jashni, M. K., Dols, I. H., Iida, Y., Boeren, S., Beenen, H. G., Mehrabi, R., ... & de Wit, P. J. (2015). Synergistic action of a metalloprotease and a serine protease from *Fusarium oxysporum* f. sp. *lycopersici* cleaves chitin-binding tomato chitinases, reduces their antifungal activity, and enhances fungal virulence. *Molecular Plant-Microbe Interactions*, 28 (9), 996-1008.

Jones, J. D., & Dangl, J. L. (2006). The plant immune system. *Nature*, 444 (7117), 323-329.

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., ... & Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics*, 30 (9), 1236-1240.

Kessens, R., Ashfield, T., Kim, S. H., & Innes, R. W. (2014). Determining the GmRIN4 requirements of the soybean disease resistance proteins Rpg1b and Rpg1r using a *Nicotiana glutinosa*-based agroinfiltration system. *PLoS One*, 9 (9), e108159.

King, E. O., Ward, M. K., & Raney, D. E. (1954). Two simple media for the demonstration of pyocyanin and fluorescin. *The Journal of laboratory and clinical medicine*, 44 (2), 301-307.

Koch, E., Ebrahim-Nesbat, F., & Hoppe, H. H. (1983). Light and electron microscopic studies on the development of soybean rust (*Phakopsora pachyrhizi* Syd.) in susceptible soybean leaves. *Journal of Phytopathology*, 106 (4), 302-320.

Koeck, M., Hardham, A. R., & Dodds, P. N. (2011). The role of effectors of biotrophic and hemibiotrophic fungi in infection. *Cellular microbiology*, 13 (12), 1849-1857.

Krombach, S., Reissmann, S., Kreibich, S., Bochen, F., & Kahmann, R. (2018). Virulence function of the *Ustilago maydis* sterol carrier protein 2. *New Phytologist*, 220 (2), 553-566.

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. *Molecular biology and evolution*, 35 (6), 1547.

Kunjeti, S. G., Iyer, G., Johnson, E., Li, E., Broglie, K. E., Rauscher, G., & Rairdan, G. J. (2016). Identification of *Phakopsora pachyrhizi* candidate effectors with virulence activity in a distantly related pathosystem. *Frontiers in Plant Science*, 7, 269.

Kupfer, D. M., Drabenstot, S. D., Buchanan, K. L., Lai, H., Zhu, H., Dyer, D. W., ... & Murphy, J. W. (2004). Introns and splicing elements of five diverse fungi. *Eukaryotic cell*, 3 (5), 1088-1100.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, *34* (18), 3094-3100.

Li, S., Smith, J. R., Ray, J. D., & Frederick, R. D. (2012). Identification of a new soybean rust resistance gene in PI 567102B. *Theoretical and Applied Genetics*, 125, 133-142.

Li, W. H., Yang, J., & Gu, X. (2005). Expression divergence between duplicate genes. *Trends in Genetics*, 21 (11), 602-607.

Link, T. I., Lang, P., Scheffler, B. E., Duke, M. V., Graham, M. A., Cooper, B., ... & Whitham, S. A. (2014). The haustorial transcriptomes of *Uromyces appendiculatus* and *Phakopsora pachyrhizi* and their candidate effector families. *Molecular Plant Pathology*, 15 (4), 379-393.

Liu, T., Song, T., Zhang, X., Yuan, H., Su, L., Li, W., ... & Dou, D. (2014). Unconventionally secreted effectors of two filamentous pathogens target plant salicylate biosynthesis. *Nature communications*, 5 (1), 4686.

Lorrain, C., Petre, B., & Duplessis, S. (2018). Show me the way: rust effector targets in heterologous plant systems. *Current opinion in microbiology*, 46, 19-25.

Lorrain, C., Gonçalves dos Santos, K. C., Germain, H., Hecker, A., & Duplessis, S. (2019). Advances in understanding obligate biotrophy in rust fungi. *New Phytologist*, 222 (3), 1190-1206.

Luderer, R., Takken, F. L., Wit, P. J. D., & Joosten, M. H. (2002). *Cladosporium fulvum* overcomes Cf‐2‐mediated resistance by producing truncated AVR2 elicitor proteins. *Molecular microbiology*, 45 (3), 875-884.

Lynch, M., & Force, A. (2000). The probability of duplicate gene preservation by subfunctionalization. *Genetics*, 154 (1), 459-473.

Ma, Z., & Michailides, T. J. (2005). Advances in understanding molecular mechanisms of fungicide resistance and molecular detection of resistant genotypes in phytopathogenic fungi. *Crop Protection*, 24 (10), 853-863.

Marguerat, S., & Bähler, J. (2010). RNA-seq: from technology to biology. *Cellular and molecular life sciences*, 67, 569-579.

Matlin, A. J., Clark, F., & Smith, C. W. (2005). Understanding alternative splicing: towards a cellular code. *Nature reviews Molecular cell biology*, 6 (5), 386-398.

Meyer, J. D., Silva, D. C., Yang, C., Pedley, K. F., Zhang, C., van de Mortel, M., ... & Graham, M. A. (2009). Identification and analyses of candidate genes for Rpp4-mediated resistance to Asian soybean rust in soybean. *Plant Physiology*, 150 (1), 295-307.

Milgroom, M. G. (2015). *Population biology of plant pathogens: genetics, ecology, and evolution*. St. Paul, MN, USA: APS Press, The American Phytopathological Society.

Morales, A. M., O'Rourke, J. A., Van De Mortel, M., Scheider, K. T., Bancroft, T. J., Borém, A., ... & Graham, M. A. (2013). Transcriptome analyses and virus induced gene silencing identify genes in the Rpp4-mediated Asian soybean rust resistance pathway. *Functional Plant Biology*, 40 (10), 1029-1047.

Muzafar, S., Sharma, R. D., Chauhan, N., & Prasad, R. (2021). Intron distribution and emerging role of alternative splicing in fungi. *FEMS Microbiology Letters*, 368 (19).

Ngou, B. P. M., Ahn, H. K., Ding, P., & Jones, J. D. (2021). Mutual potentiation of plant immunity by cell-surface and intracellular receptors. *Nature*, 592 (7852), 110-115.

Olivieri, F., Eugenia Zanetti, M., Oliva, C. R., Covarrubias, A. A., & Casalongué, C. A. (2002). Characterization of an extracellular serine protease of *Fusarium eumartii* and its action on pathogenesis-related proteins. *European Journal of Plant Pathology*, 108, 63-72.

Ortiz, D., Chen, J., Outram, M. A., Saur, I. M., Upadhyaya, N. M., Mago, R., ... & Dodds, P. N. (2022). The stem rust effector protein AvrSr50 escapes Sr50 recognition by a substitution in a single surface‐exposed residue. *New Phytologist*, 234 (2), 592-606.

Pedley, K. F., Pandey, A. K., Ruck, A., Lincoln, L. M., Whitham, S. A., & Graham, M. A. (2019). *Rpp1* encodes a ULP1-NBS-LRR protein that controls immunity to *Phakopsora pachyrhizi* in soybean. *Molecular Plant-Microbe Interactions*, 32 (1), 120-133.

Persoons, A., Hayden, K. J., Fabre, B., Frey, P., De Mita, S., Tellier, A., & Halkett, F. (2017). The escalatory Red Queen: Population extinction and replacement following arms race dynamics in poplar rust. *Molecular Ecology*, 26 (7), 1902-1918.

Petit-Houdenot, Y., & Fudal, I. (2017). Complex interactions between fungal avirulence genes and their corresponding plant resistance genes and consequences for disease resistance management. *Frontiers in plant science*, 8, 1072.

Porto, B. N., Caixeta, E. T., Mathioni, S. M., Vidigal, P. M. P., Zambolim, L., Zambolim, E. M., ... & Resende, M. L. V. D. (2019). Genome sequencing and transcript analysis of *Hemileia vastatrix* reveal expression dynamics of candidate effectors dependent on host compatibility. *PLoS One*, 14 (4), e0215598.

Prasad, P., Savadi, S., Bhardwaj, S. C., Gangwar, O. P., & Kumar, S. (2019). Rust pathogen effectors: perspectives in resistance breeding. *Planta*, 250, 1-22.

Pruitt, R. N., Locci, F., Wanke, F., Zhang, L., Saile, S. C., Joe, A., ... & Nürnberger, T. (2021). The EDS1–PAD4–ADR1 node mediates Arabidopsis pattern-triggered immunity. *Nature*, 598 (7881), 495-499.

Qi, M., Link, T. I., Müller, M., Hirschburger, D., Pudake, R. N., Pedley, K. F., ... & Whitham, S. A. (2016). A small cysteine-rich protein from the Asian soybean rust fungus, *Phakopsora pachyrhizi*, suppresses plant immunity. *PLoS Pathogens*, 12 (9), e1005827.

Rabouille, C. (2017). Pathways of unconventional protein secretion. *Trends in cell biology*, 27 (3), 230-240.

Roach, M. J., Schmidt, S. A., & Borneman, A. R. (2018). Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics*, 19 (1), 1-10.

Rocafort, M., Fudal, I., & Mesarich, C. H. (2020). Apoplastic effector proteins of plant-associated fungi and oomycetes. *Current opinion in plant biology*, 56, 9-19.

Rovenich, H., Boshoven, J. C., & Thomma, B. P. (2014). Filamentous pathogen effector functions: of pathogens, hosts and microbiomes. *Current opinion in plant biology*, 20, 96-103.

Saito, C., Ueda, T., Abe, H., Wada, Y., Kuroiwa, T., Hisada, A., ... & Nakano, A. (2002). A complex and mobile structure forms a distinct subregion within the continuous vacuolar membrane in young cotyledons of *Arabidopsis*. *The Plant Journal*, 29, 245-255.

Salcedo, A., Rutter, W., Wang, S., Akhunova, A., Bolus, S., Chao, S., ... & Akhunov, E. (2017). Variation in the *AvrSr35* gene determines *Sr35* resistance against wheat stem rust race Ug99. *Science*, 358 (6370), 1604-1606.

Sambrook, J., Fritsch, E. F., & Maniatis, T. (1989). *Molecular cloning: a laboratory manual* (No. Ed. 2). Cold spring harbor laboratory press.

Schwessinger, B., & Zipfel, C. (2008). News from the frontline: recent insights into PAMP-triggered immunity in plants. *Current opinion in plant biology*, 11 (4), 389-395.

Schwessinger, B., & Rathjen, J. P. (2017). Extraction of high molecular weight DNA from fungal rust spores for long read sequencing. *Wheat Rust Diseases: Methods and Protocols*, 49-57.

Silva, D. C., Yamanaka, N., Brogin, R. L., Arias, C. A., Nepomuceno, A. L., Di Mauro, A. O., ... & Abdelnoor, R. V. (2008). Molecular mapping of two loci that confer resistance to Asian rust in soybean. *Theoretical and Applied Genetics*, 117, 57-63.

Sohn, K. H., Lei, R., Nemri, A., & Jones, J. D. (2007). The downy mildew effector proteins ATR1 and ATR13 promote disease susceptibility in *Arabidopsis thaliana*. *The Plant Cell*, 19 (12), 4077-4090.

Sonah, H., Deshmukh, R. K., & Bélanger, R. R. (2016). Computational prediction of effector proteins in fungi: opportunities and challenges. *Frontiers in plant science*, 7, 126.

Smukowski Heil, C. (2023). Loss of Heterozygosity and Its Importance in Evolution. *Journal of Molecular Evolution*, 1-9.

Sperschneider, J., Gardiner, D. M., Dodds, P. N., Tini, F., Covarelli, L., Singh, K. B., ... & Taylor, J. M. (2016). EffectorP: predicting fungal effector proteins from secretomes using machine learning. *New Phytologist*, 210 (2), 743-761.

Sperschneider, J., Dodds, P. N., Taylor, J. M., & Duplessis, S. (2017). Computational methods for predicting effectors in rust pathogens. *Wheat rust diseases: methods and protocols*, 73-83.

Stephenson, S. A., Hatfield, J., Rusu, A. G., Maclean, D. J., & Manners, J. M. (2000). *CgDN3*: an essential pathogenicity gene of *Colletotrichum gloeosporioides* necessary to avert a hypersensitive-like response in the host *Stylosanthes guianensis*. *Molecular plant-microbe interactions*, 13 (9), 929-941.

Stergiopoulos, I., & de Wit, P. J. (2009). Fungal effector proteins. *Annual review of phytopathology*, 47, 233-263.

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., & Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, 20, 467-484.

Thomas, W. J., Thireault, C. A., Kimbrel, J. A., & Chang, J. H. (2009). Recombineering and stable integration of the *Pseudomonas syringae* pv. *syringae* 61 *hrp/hrc* cluster into the genome of the soil bacterium *Pseudomonas fluorescens* Pf0‐1. *The Plant Journal*, 60 (5), 919-928.

Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics*, 14 (2), 178-192.

Tobias, P. A., Schwessinger, B., Deng, C. H., Wu, C., Dong, C., Sperschneider, J., ... & Park, R. F. (2021). *Austropuccinia psidii*, causing myrtle rust, has a gigabase-sized genome shaped by transposable elements. *G3: Genes, Genomes, Genetics*, 11 (3).

Upadhyaya, N. M., Mago, R., Panwar, V., Hewitt, T., Luo, M., Chen, J., ... & Dodds, P. N. (2021). Genomics accelerated isolation of a new stem rust avirulence gene–wheat resistance gene pair. *Nature Plants*, 7 (9), 1220-1228.

van den Hooven, H. W., van den Burg, H. A., Vossen, P., Boeren, S., de Wit, P. J., & Vervoort, J. (2001). Disulfide bond structure of the AVR9 elicitor of the fungal tomato pathogen *Cladosporium fulvum*: evidence for a cystine knot. *Biochemistry*, 40, 3458-3466.

van der Auwera, G. A., & O'Connor, B. D. (2020). *Genomics in the cloud: using Docker, GATK, and WDL in Terra*. O'Reilly Media.

Ve, T., Williams, S. J., Catanzariti, A. M., Rafiqi, M., Rahman, M., Ellis, J. G., ... & Kobe, B. (2013). Structures of the flax-rust effector AvrM reveal insights into the molecular basis of plant-cell entry and effector-triggered immunity. *Proceedings of the National Academy of Sciences*, 110 (43), 17594-17599.

Wang, C. I. A., Guncar, G., Forwood, J. K., Teh, T., Catanzariti, A. M., Lawrence, G. J., ... & Kobe, B. (2007). Crystal structures of flax rust avirulence proteins AvrL567-A and-D reveal details of the structural basis for flax disease resistance specificity. *The Plant Cell,* 19 (9), 2898-2912.

Wapinski, I., Pfeffer, A., Friedman, N., & Regev, A. (2007). Natural history and evolutionary principles of gene duplication in fungi. *Nature*, 449(7158), 54-61.

Webb, C. A., & Fellers, J. P. (2006). Cereal rust fungi genomics and the pursuit of virulence and avirulence factors. *FEMS microbiology letters*, 264 (1), 1-7.

Wu, W., Nemri, A., Blackman, L. M., Catanzariti, A. M., Sperschneider, J., Lawrence, G. J., ... & Hardham, A. R. (2019). Flax rust infection transcriptomics reveals a transcriptional profile that may be indicative for rust *Avr* genes. *Plos one*, 14 (12).

Yuan, M., Jiang, Z., Bi, G., Nomura, K., Liu, M., Wang, Y., ... & Xin, X. F. (2021). Pattern-recognition receptors are required for NLR-mediated plant immunity. *Nature*, 592 (7852), 105-109.

Zambolim, L., Reis, E. M., Guerra, W. D., Juliatti, F. C., & Menten, J. O. M. (2022). Integrated Management of Asian Soybean Rust. *European Journal of Applied Sciences*. Vol, 10 (2).

**SUPPLEMENTARY MATERIAL**

**Supplementary Table 1**: *P. pachyrhizi* isolates sequenced and analyzed in this study and their virulence phenotype on soybean genotypes containing *Rpp1b* or *Rpp5*.

| Isolate | Collected in | Virulence phenotype | |
|---|---|---|---|
| | | Rpp1b | Rpp5 |
| PHPA_01 | Mato Grosso (Brazil) | **Virulent** | Avirulent |
| PHPA_02 | Mato Grosso (Brazil) | Avirulent | Avirulent |
| PHPA_03 | Mato Grosso (Brazil) | **Virulent** | Avirulent |
| PHPA_04 | Mato Grosso (Brazil) | Avirulent | Avirulent |
| PHPA_05 | Mato Grosso (Brazil) | **Virulent** | Avirulent |
| PHPA_06 | Mato Grosso (Brazil) | **Virulent** | Avirulent |
| PHPA_07 | Mato Grosso (Brazil) | Avirulent | Avirulent |
| PHPA_08 | Japan | Avirulent | Avirulent |
| PHPA_09 | Japan | **Virulent** | Avirulent |
| PHPA_10 | Japan | Avirulent | Avirulent |
| PHPA_11 | Paraná (Brazil) | Avirulent | Avirulent |
| PHPA_12 | PHPA_11 derived mutant | Avirulent | **Virulent** |
| PHPA_13 | Paraná (Brazil) | Avirulent | Avirulent |
| PHPA_17 | Paraguay | Avirulent | Avirulent |
| PHPA_18 | Paraguay | Avirulent | Avirulent |
| PHPA_19 | Paraguay | Avirulent | Avirulent |
| PHPA_20 | Paraguay | Avirulent | Avirulent |
| PHPA_21 | Minas Gerais (Brazil) | Avirulent | Avirulent |
| PHPA_22 | Mato Grosso (Brazil) | **Virulent** | Avirulent |
| PHPA_23 | Mato Grosso (Brazil) | Avirulent | Avirulent |
| PHPA_24 | Mato Grosso (Brazil) | **Virulent** | Avirulent |
| PHPA_25 | PHPA_24 derived mutant | Avirulent | Avirulent |
| PHPA_26 | PHPA_23 derived mutant | **Virulent** | Avirulent |
| PHPA_28 | Japan | Avirulent | **Virulent** |
| PHPA_29 | Minas Gerais (Brazil) | Avirulent | Avirulent |
| PHPA_30 | Japan | **Virulent** | Avirulent |
| PHPA_31 | Japan | **Virulent** | Avirulent |
| PHPA_32 | Rio Grande do Sul (Brazil) | Avirulent | Avirulent |
| PHPA_33 | Rio Grande do Sul (Brazil) | Avirulent | Avirulent |
| PHPA_34 | Bahia (Brazil) | Avirulent | Avirulent |
| PHPA_35 | Santa Catarina (Brazil) | Avirulent | Avirulent |
| PHPA_36 | Paraná (Brazil) | Avirulent | Avirulent |
| PHPA_37 | Paraguay | Avirulent | Avirulent |
| PHPA_38 | Paraguay | Avirulent | Avirulent |
| PHPA_39 | Paraguay | Avirulent | Avirulent |
| PHPA_40 | Paraguay | Avirulent | Avirulent |
| 71G2 | Mato Grosso (Brazil) | Avirulent | Avirulent |
| 71G3 | 71G2 derived mutant | **Virulent** | Avirulent |

**Supplementary data:**

>cAvr1-PHPA79-254_1151Kb-Ed-4444549_allele-1

MKRSYTYLFFSVVFGAFSLGYGFVPSEPSELSEIEGQRNITERSLEPPIDLTKTGYYTKGCYKSD
QKPVLVEDCLQLMENMKNNQDLVECGPGCCKYWPYKSCSVFLAVSSSAKRNYQISAFSLGDT
ISGTMRDCQQQNSEIFTGGWYSFYTPENVEWESQNPNSPDHPPVVTTVSATPN*

>cAvr1-PHPA79-254_1151Kb-Ed-4444549_allele-2

MKRSYTYLFFSVVFGAFSLGYGFVPSEPSELSEIEGQRNITERSLEPPIDLTKTGYYTKGCYKSD
QKPVLVEDCLQLMENIKNNQDLVECGPGCCKYWPYKSCSVFLAVSSSAKRNYQISAFSLGDTI
SGTMRDCQQQNSEIFTGGWYSFYTPENVEWESQNPNSPDHPPVVTSVSATPN*

>cAvr1-PHPA79-254_1125Kb-Ed-3684774

MKRSYTYLFFSVVFGAFSLGYGFVPSEPSELSEIEGQRNITERSLEPPIDLTKTGYYTKGCYKSD
QKPVLVEDCLQLMENIKNNQDLVECGPGCCKYLPYKSCSVFLYVSSSAKRNYQISAFSLGDTIS
GTMRNCQQQNSEIYPGGWYSFYTPENVEWESQNPNSPDYPPVVTVVSSTPN*

>cAvr1-PHPA79-221_967Kb-Ed-3375891_Allele-1

MKRSYTYFFFSVVFGAISSGYGFLPSGSSELSEIEGQRNITERSLEPPIDLTKTGYYMKGCYKSD
QKPVLVEDCLQLMENIKNNQDPVECGPGCNRYLPYKSCAVFFMVSSSAKRNYQIAEFSLGDTI
SGTMRNCQEQNSESYPGGWYSFYTPENVEWESQNPNSPDYPPVITAVTATPN*

>cAvr1-PHPA79-221_967Kb-Ed-3375891_Allele-2

MKRSYTYFFFSVVFGAISSGYGFLPSGSSELSEIEGQRNITERSLEPPIDLTKTGYYMKGCYKSD
QKPVLVEDCLQLMENIKNNQDPVECGPGCNRYLPYKSCAVFFMVSSSAKRNYQIAEFSLSDTI
SGTMRNCQEQNSESYPGGWYSFYTPENVEWESQNPNSPDYPPVITAVTATPN*

>cAvr1-60_267Kb-492078

MKTSVSKPRVKLISLVNLRRPTTTTTTTTATSKIQLKQLQHSTPTILSTQSIPTITTADNTKTCCYS
YPSTSTSLITNHQLPSLIINSCNRIQRSSSSIRSTTTIILPNLINTNNQLRLYYSSNNSSSNSNSNSN
NIQSINQLNRLNQYQNHHLLHHHPILRNIFTQTNFNHSILTFFEPSIHPNLQQHPSSSHHQNFINN
SSIDCVCYNQPQPKPNSSPSNCHQNHSNNSSSPRKSSNTSNSSPSSTLLPQSSQSKSQNPNHQSS
SSSSSSSSNNKFTGSINSNLVSLTQFNHLNIYPTSADEPNDSVGPSIRSVLPRKTFYLFRNGASGIP
KSSPTSSPSLFLMSSSSSTTTPTTSTQSTHGTDGYGDRVTLSSSSRKDQSINRSSPSCSTYRAVNS
VQIGEDSYFLRNDSLGVADGVGGWSGKPGANAGLFSSKLMNHCYNEISRYENTEDDRFQSYN
DIDPVEILQRAFEMSIYESKDEGILGSSTALVAILRNDELRIANIGDCCCSIIRGNDYIFRTEEQQH
SFNYPVQIGTNSKSVPIRDAQRYKVKVQKDDVVILSSDGLVDNLFDEDILEEVLKFTKPKNRLS
LRSASGTSEIKSLEEEGMRLKGDGVGSSSSSSNRGNENGRTKMIGSVPELISKSLCIRAKTVIDD
QQAITSPFAQRASEEGIHYVGGKNDDISVLVAIVGALES*

>cAvr5-138_913Kb-Allele-1

MFKTELLKQPSIAKDIKREKSSRLLSGTLKKMS*

>cAvr5-138_913Kb-Allele-2

MFKTELLKQPSIAKVIKREKSSSLLSGALKKMS*

>cAvr5-47_1128Kb-3726186_allele-1

MYYFSFFYIFLLSSLMRYLVIGSSSSDECDSGKISPDTWKEKKIDEQLANFPSGKKLSLEAFASF
VGYDGFRCGIGEDCLAGQICSPVKAPYWQVLVAAQEWNFYINRVYQAFSFSVATVKVPATIMT
ADVCLGIYVLTIGLSALLIMPSPMTAVVTPFFRTSLLGGVATYISLPELATYYSSRPQKSEFLSTIS
SEISEWENETHQRINKNLNEIINDKPIVSSGGLNEILANGTFFTNSSMFDTTHIFNNYRKVIQIRS
LVSILRSKSDCDGPGKDGMDIKENHLSWCDSSKKMTKIVFAHGSKIKRKIKGANLIAYKYGYS
TKFLTISSLECQKKYGVGADLYKNGHTIGKKKTRNFSCLSWSRIFDNLRINTKR*

>cAvr5-134_549Kb-6548230_allele-1

MPKYSNFKMFSNVLVSMWIFLMIFSLLHSASASLERVVELTEASKASDETKNSLEVEKKTISAS
HPQIQQENSNLEMSLPLINGPNKNANSNKLNREEKQSTELKNRNTSKVARHPPTSQQTPSLHE
TNPIGDPNLSNLHHNSLPSYGTQTADPRSLLVQPILIPDVNYVGPVVVPTVYDLFEYGYRQSYV
PLAQIIPSTEHKISGLENFGQSKGTGVYHGHFEKDETPYIPPDQLKPSKAESRLLYKQISSEKFVP
SHGENIPKFRIYKPSLKEKSEPSSNNQVSESSKEIDSNLENEKPTQYLKVENNIYDNTHSMEKK
KFDFMLEKSNKQPEPKGLKAISISKENKSTNQFSIHDSNSSLADDSRENIKGNIKKIDSGLKDAI
FYDDYPIQIGSIKIPASRWRPQNSAESRFHAAKPAFHEIFDSSSKNIPMDSIPIPSEKEKNKSLSEN
KIGETSKKTNIDEDKKKSNQTPPVPKKLNYNTHAKKNKKFDEVDEKLDEQPKNQGLKTTLTT
KENKSENLSVTQFNTHDMHSVPIGDSDQNLKRENNKINTELKDARFYGDYSIQGGLSKIPTSH
RRHQNKAESQFHAVQPAQSNKFHSSNRKNSLINRMPSALQKEKILALTKNQVGDSSNETDIDK
DKEKLNELLPAKKNENYNTHLEEKNKVKDMDEELNEQPKNQVLKSASISKNKKKKKEFVNQ
PNTHDSHSAFVDAAIQGLKGKNKTIFSQNKEDMIGEDSLNQKTESPELSNKKDNFLPLQSNKK
NIRANYFSNRSLLHNQNFKEKQLIQKFDDNESNILKVKETLINTADDFKDVNEIYTEKASVDSS
KLSTKYNSSKSKSPLNTDLSIPLSVLYEDNSKNSEILETSEKLKETFSKSQFNLKDSELGKKADQ
GKANGSNAFEDNFPRNFIESKPKSVKNSKKGSKHKGGLGNSKKKLNKNHSESELKSDTDTKK
PSDVQNLVSSSIVNKHFVNKLIDENVTKPEINSKHNTNLGQKIFNAIAESSVPKETTNSSFVKGN
NKSFPKIKFVSLDSNKKAEQEISKLFYDATILLGDKRAIDLVLQSFPKVVNSFIITLIQKILVVKE
DSDKEAIEDVFHTIEHAASEEPQNLHTAIKWLRNSISDSEELIEKEKYFNKMFFNWLVRKKNQE
NSGKNFIDKLQEIVKSHPLYYGN*

>cAvr5-310_66Kb-692760_allele-1

MSWSSLHTTCLAGFLSFLVLTQVFSAPNQERSPDTPALGIGSASSSSQLTPSGALPQDVAQSQSP
ALGRGNLTQASSNTPAQDSHLTSPASNDPSHGPALINLTQVQEPSISSVAGSTAQSNQSHAPPPST
ALAQSTSSNNPTPLATSVSSPISFLKNLTNESESKAKNNLSVTFPVTSFNGFNESNVGVSIKVSG
VDDAISLSDNKTHDPVPVIEVGLFHFNSASSKVFYVPGNKIPFVTWNVGPSYAGMLPVSPAVIG
GNITDDNHLGVSLNNPTTNVSVGAGGHEVPVGSKPPGIDINVGTHPSRLRLRSSSSDPTSTTIPL
TSGSQRSNSTTEQIQDSSKKIFFWFFPATAVEGTNKLTLWLMDKFFKYLIRRG*

The other secretome gene and protein sequences an the Blastp results are available at the following folder:

https://drive.google.com/file/d/1h8dsBJImHKeS2kUDZhIkgne-GXlH066Z/view?usp=sharing

# CHAPTER 2

## GENETIC AND GENOMIC STRUCTURE OF THE MATING-TYPE LOCI OF SOYBEAN-ASSOCIATED *Phakopsora* SPECIES

## ABSTRACT

Mating compatibility in fungi is controlled by mating-type loci and it could be necessary for syngamy and other sexual reproduction processes. Most of the rust pathogens have a tetrapolar system, coordinated by two unlinked mating loci encoding pheromone precursor peptides, receptors (Mfa and STE3.2), and homeodomain transcription factors (HD1 and HD2). Few representatives of Pucciniales had their mating-type genes studied, not including any species of the genus *Phakopsora*. *P. pachyrhizi,* the most important fungal pathogen of soybean, never had its complete sexual cycle reported nor its mating type system studied. Characterization of the *Phakopsora* mating-type system can improve the knowledge of the evolution of the Pucciniales mating system and the sexual reproduction of *P. pachyrhizi*. In this study, the main objective was to use whole genome sequencing data of *P. pachyrhizi*, *P. meibomiae*, and *Phakopsora* sp. to characterize the genetic and genomic structure of the mating type loci of these *Phakopsora* species infecting Fabaceae and to infer if *P. pachyrhizi* contains possible functional mating-type genes that enable the sexual cycle of this species. Using mating-type protein sequences from other rust fungi and RNA-seq data from *P. pachyrhizi*, the mating-type genes encoding the STE3.2, Mfa, and HD proteins in the three species were annotated. Protein sequences were used for the phylogeny of STE3.2 and HD and the similarity among the Mfa peptides was determined. The conserved domains and the 3D structure for the receptor proteins were predicted and compared with homologous proteins from other rust fungal species. A probable tetrapolar system was identified in all *Phakopsora* species, where the *HD* and *Ste3.2/mfa* loci were not physically linked. When compared with other rust species, multiple paralogs of *mfa* genes and longer physical distances between *mfa* and *Ste3.2* genes were characteristics of the studied *Phakopsora* species. The *Phakopsora* species differ in their mating type protein sequences and in the number and organization of the genes in the *Ste/mfa* loci. *P. pachyrhizi Ste3.2-2* was predicted to encode an atypical seven-transmembrane receptor protein with a loss of transmembrane domains and a less compact structure. In conclusion, it was demonstrated that the *Phakopsora* species analyzed contains a possible tetrapolar system with a conserved structure for the *HD* loci and a variable structure at the *Ste3.2/mfa* loci.


Keywords: Sexual reproduction. Pheromone. GPCR. Asian soybean rust.

## 1. INTRODUCTION

Sexual reproduction occurs in different species of fungi and usually involves syngamy, karyogamy, and meiosis. It is an important process for species evolution that increases recombination frequency and usually promotes higher genetic diversity, faster adaptation to environmental changes, more efficient selection of beneficial mutations, and purge of deleterious mutations, resulting in resistant propagules that are essential for the fungi to survive in periods of unfavorable environmental conditions (McDonald & Linde, 2002; Ni et al., 2011; Heitman et al., 2013). The sexual cycle could be explored in mycology and genetics for breeding (Esser, 1971), gene mapping (Foulongne-Oriol, 2012) and to understand the population biology and structure (Paoletti et al., 2005; Rydholm et al., 2006).

Mating compatibility in fungi is controlled by mating-type (*MAT*) loci, a set of genes involved in syngamy and advancing to karyogamy between individuals (Ni et al., 2011; Coelho et al., 2017). *MAT* loci vary widely in their organization among species due to different evolutionary histories. While the ascomycetes mating type is controlled by one locus (unifactorial), basidiomycetes mating usually is controlled by two loci (bifactorial), generating bipolar or tetrapolar arrangements if these loci are or are not physically linked, respectively (Kües et al., 2011; Coelho et al., 2017). In both Dykaria phyla, there are species in which haploid cells could be universally compatible (homothallic), making possible reproduction even between clonemates, and species that need individuals with different mating types for mating (heterothallic), resulting in sexual reproduction and higher diversity (Coelho et al., 2017).

Basidiomycetes mating-type system is organized in two genetic mating loci, encoding: 1) the pheromone receptor (STE3.2) of the seven transmembrane receptor family (7TMR) and the pheromone precursor peptide (mfa) at locus *Ste3/Mfa* (also called *P/R*); and 2) the homeodomain-type transcription factors (HD1 and HD2) at locus *HD* (also called *bE*/*bW*) (Kües et al., 2011) (Figure 1). Both loci are essential for the mating process. *Mfa* genes encode pheromone precursor peptides which are post-translationally modified and processed in many pheromone molecules with 9-14 amino acids (Bakkeren et al., 2008). *Ste3* gene encodes a 7TMR coupled to G-proteins inside the cell. The diffusible pheromones bind to the cognate receptor and activate downstream signal transduction mediated by mitogen-activated protein kinase (MAPK). The *HD* locus encodes two unrelated HD transcription factors that can form a heterodimeric complex. The heterodimeric HD1-HD2 transcription factor regulates the expression of genes acting in dikaryon development and sexual reproduction

(Bakkeren et al., 2008). Successful mating only occurs when opposite mating types are involved, leading to the interaction between pheromones with their cognate pheromone receptors and the formation of the heterodimer using alleles present in different mating types (Bakkeren et al., 2008) (Figure 1). When the two mating-type loci are physically unlinked four mating types could be generated by meiosis resulting in the tetrapolar breeding system. In species that carry these loci linked or when one of these loci loses its function the tetrapolar system turns into a bipolar (Coelho et al., 2017).

Pucciniales (Pucciniomycotina) is one of the most important orders of Basidiomycetes, causing plant diseases known as rusts. Fungi causing rust on plants are biotrophic pathogens with a complex life cycle that could include more than one host (heteroecious) and up to five different kinds of spores to complete their cycle (macrocyclic species) (Duplessis et al., 2021). Historically, the usually complex genome of rust fungi, including its large size and high percentage of repetitive elements, had limited their genome assembly and further genomic analysis (Duplessis et al., 2014; Loehrer et al., 2014). Using short-read DNA sequencing, a few rust pathogen species, that have relatively small genomes compared with other members of the order, were sequenced and had their genomes assembled (Duplessis et al., 2014). Recent advances in sequencing technology with long-read DNA sequencing [Single Molecule, Real-Time - SMRT (Pacbio) and Nanopore Sequencing (Oxford)] made it possible new genome assemblies of rust species enabling the use of genomic data in studies with these pathogens (Gupta et al., 2023).

Using the available genome assemblies, the mating-type genes were annotated in the following rust species: *Melampsora larici-populina* (Duplessis et al., 2011)*, Cronartium quercuum-fusiforme* (Pendleton et al., 2014), *P. striiformis* f.sp. *tritici*, *P. triticina* and *P. graminis* f.sp. *tritici* (Cuomo et al., 2017) and *Austropuccinia psidii* (Ferrarezi et al., 2022). For now, studies of mating-type systems in rust fungi have identified: a) the *HD* locus with two genes *(HD1* and *HD2*) and two alleles, present in all studied species; b) three *pheromone receptor genes* (*Ste3.2-1, Ste3.2-2* and *Ste3.2-3*) in most of the species, and four *Ste3.2* (*1-4*) in *M. larici-populina* (Duplessis et al., 2011); and c) one *mfa2* gene in *P. striiformis* f.sp. *tritici*, *P. triticina* and *P. graminis* f.sp. *tritici*, one additional putative *mfa3* gene in *P. graminis* f.sp. *tritici* (Cuomo et al., 2017) and ten putative *mfa genes* in *M. larici-populina* (Duplessis et al., 2011). Tetrapolar breeding was supposed for all those species (Cuomo et al., 2017; Coelho et al., 2017). Recently, high-quality chromosome-level assemblies of the genomes of *P. triticina* and *P. graminis* f.sp *tritici* confirmed the tetrapolar system since the *HD* and *Ste*/*mfa* loci are located in distinct chromosomes in these species. The pseudo-chromosome

assemblies also supported that *Ste3.2-2* and *Ste3.2-3* are haplotype-specific (present in just one of the two nuclei), while the gene *Ste3.2-1* has two alleles, it is present in both nuclei and probably is not related to the mating-type compatibility system (Li et al., 2019; Wu et al., 2021). Unfortunately, the mating-type system was not deeply studied in all these species, but the difference in the number of genes indicates that the mating-type system could vary among rust species.

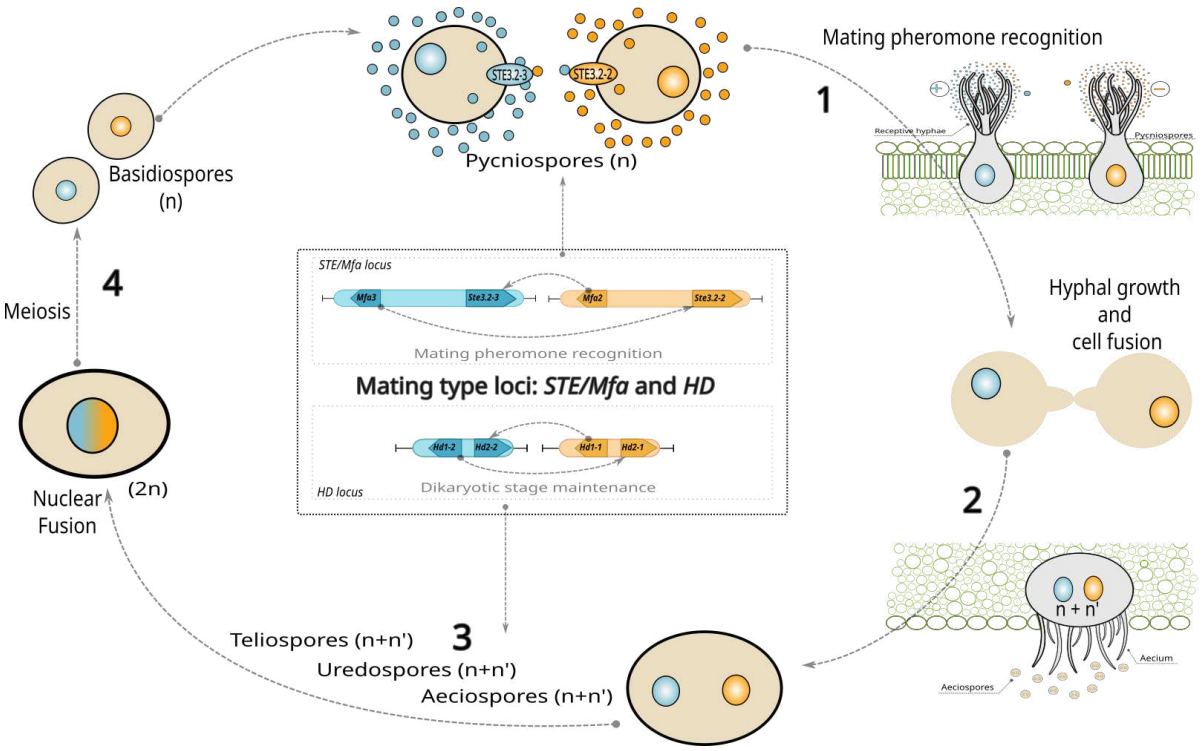



**Figure 1**: Schematic representation of the mating type process in the macrocycle of a heteroecious rust fungus. Haploid (n) individuals during the pycnia stage recognize the pheromone from a different mating variant due to its interaction with the pheromone receptor (1). The mating recognition induces hyphal growth and cell fusion (plasmogamy) between the individuals, forming a dikaryon in the aecium stage (2). The dikaryotic phase is maintained with the presence of the HD transcription factors HD-1 and HD-2 encoded from nuclei of both mating variants (3). During the germination of the teliospores occurs nuclear fusion (karyogamy) and meiosis that results in haploid basidiospores (4). The mating type loci *STE/Mfa* and *HD* are represented inside the circular cycle. The mating locus *STE/Mfa* encodes the pheromone precursor (Mfa) and the pheromone receptor (STE3.2) associated with the mating recognition in the pycnia stage and the *HD* locus encodes two bi-allelic homeodomain-type transcription factors (HD1 and HD2) that interact (HD1-HD2) using the alleles encoded from opposite mating types to form a heterodimeric complex that maintains the dikaryon. Blue or orange color indicates the two different nuclei and the mating types associated with each nucleus. The mating only occurs with the interaction of proteins of opposite mating types encoded in both locus *STE/Mfa* and *HD* (indicated by the intern arrows in the box with the mating type loci).

Despite the importance of the *Phakopsora* as plant pathogens, the mating-type system was never studied in this genus. Two *Phakopsora* species are reported causing rust disease on

soybeans: *P. meibomiae* and *P. pachyrhizi*. The first one is natural from the Americas and causes American soybean rust, with a minor economic impact on soybean production (Goellner et al., 2010). The second is an extremely aggressive pathogen on soybean that demands a high volume of fungicides for its management, causing a negative environmental and economic impact on Brazilian soybean fields (Yorinori et al., 2005; Hartman et al., 2015). Both species reproduce using urediniospores and no sexual reproduction has been reported yet (Yamaoka, 2014). However, the possibility of the sexual cycle occurs in some specific conditions is not discarded and holds the attention of researchers, because the mating could increase the genotypic variability of the population and turn the disease management more difficult with the emergence of new genotypes more aggressive, virulent to multiple soybean cultivars and resistant to different fungicides (McDonald & Linde, 2002).

Understanding the mating-type system can bring answers about the possibility of having a sexual cycle in *P. pachyrhizi* and *P. meibomiae* and also clarify the evolution of the mating system in some *Phakopsora* species where the aecium stage has not been found (Ono et al., 1992; Twizeyimana et al., 2011). If sexual reproduction is possible in some scenarios, disease management must be adequate to prevent the conditions that enable mating.

The genome analysis of *P. pachyrhizi* and two other *Phakopsora* species (*P. meibomiae* and *Phakopsora* sp.), associated with rust diseases on soybean, can bring light about the unknown mating-type system in these species and also be used to identify the components necessary for sexual reproduction in *Phakopsora* species for which no aecium stage has been reported yet. Using the recently available genome assemblies of *P. pachyrhizi*, *Phakopsora* sp., and *P. meibomiae* isolates infecting soybean, the genomic and genetic structure of the mating-type loci in *Phakopsora* spp. was reported, the mating type genes were annotated and structural variations, that could be associated with the absence or unreported sexual cycle in *Phakopsora* species infecting soybeans, were predicted.

## 2. MATERIALS AND METHODS

### 2.1 Biological material

Monolesional *P. pachyrhizi* Syd. & P. Syd. isolate 71G2 was obtained from single lesions of naturally infected soybean (*Glycine max* L. Merrill) from Mato Grosso state, and *P. meibomiae* Arthur isolate MG19.3.5 was collected from *Phaseolus lunatus* in Espírito Santo state and *Phakopsora sp.* isolate MG19.8.2 (unreported species) was collect from *Lablab purpureus* in Ipiranga, Minas Gerais. One single lesion from each sample was detached and

reinoculated on healthy detached leaves to purify the isolate and reproduce the disease symptoms. Uredospores were collected from the inoculated leaves and used to multiply the isolate on healthy plants.

Soybean, *P. lunatus,* and *L. purpureus* plants used in the spore multiplication were grown in 1L plastic pots in growth chambers at 22°C, with a photoperiod of 12/12 (light/dark). Thirty days old plants were inoculated with spore suspensions with $10^5$ uredospore $mL^{-1}$. The suspension was sprayed on leaves and the inoculated plants were kept in dew chambers at 25°C for 24h, then they were moved again to the growth chambers at 22°C. Uredospores were harvested with a spore collector 15 days after the inoculation. Collected uredospores were dehydrated in silica gel for 24h and preserved at -80°C.

**2.2 DNA extraction and sequencing.**

The high molecular weight (HMW) genomic DNA from isolates 71G2, MG19.3.5, and MG19.8.2 was extracted from spores using a modified CTAB protocol (Schewessinger & Rathjen, 2017). A 20-Kb and 40 Kb PacBio SMRTbell libraries were prepared by BGI (https://www.bgi.com) with 20 and 40-Kb Blue Pippin size selection, being performed prior to sequencing on a PacBio Sequel II system (Pacific Biosciences, Menlo Park, CA). For isolates MG19.3.5 and MG19.8.2, sequence data collection was standardized to 30 hours to allow ample time for multiple pass sequencing around SMRTbell template molecules of 10–25 Kb which yields high-quality circular consensus sequencing (HiFi reads; Wenger et al., 2019) results. Raw base-called data was moved from the sequencing instrument and imported into SMRTLink to generate HiFi reads using the CCS algorithm which processed the raw data and generated the HiFi fastq files with the following settings: minimum pass 3, minimum predicted RQ 20.

**2.3 Genome assembly**

FALCON and FALCON-Unzip algorithms (https://github.com/PacificBiosciences/FALCON/) were used to assemble the PacBio SMRT long-read sequencing data from 71G2 isolate into highly accurate, contiguous, and correctly phased diploid genomes. HiFi reads from MG19.3.5 and MG19.8.2 were assembled using hifiasm (Cheng et al., 2021) with default parameters.

BUSCO version 5.2.2 (Simão et al., 2015) was used to confirm the completeness of the genome using the basidiomycetes database (basidiomycota_odb10) and *Ustilago maydis* as Augustus parameter species. The genome of isolate PPUFV02, available at Mycocosm

(https://mycocosm.jgi.doe.gov/PpacPPUFV02/PpacPPUFV02.info.html), also was used in our mating-type analysis to better characterize the system.

**2.4 Identification and sequence comparison of mating-type genes**

The protein sequences of STE3-like pheromone receptors of rust species (STE3.2), pheromone precursors (Mfa), and homeodomain-containing transcription factors (HD1 and HD2) from *P. striiformis* f. sp. *tritici* and *M. larici-populina* (Cuomo et al., 2018; Duplessis et al., 2011) were used to search for *P. pachyrhizi* genes. For this, the NCBI blast software algorithms implemented for local analysis (Morgulis et al., 2008) were used to search the PPUFV02 assembled genome. Protein sequences were manually curated associating the predicted protein with RNA-seq data (from *P. pachyrhizi* PPUFV02 isolate infecting soybeans and *in vitro* germinated spores) that was mapped on PPUFV02 genome using Minimap2 (Li, H., 2018) [parameters -ax]. The putative pheromone precursor gene *mfa3* was identified manually with Integrative Genome Viewer - IGV (Thorvaldsdóttir et al., 2013) by searching putative expressed open reading frames (ORFs) linked to the predicted *Ste3.2-3* gene, containing the following features: a) start codon (methionine); b) stop codon; c) tandem repeats and d) CAAX domain at N-terminal portion. The putative Mfa peptides were aligned against other described Mfa of rusts using Mega X (Kumat et al., 2018).

All mating-type genes (*Ste3, Mfa, HD-1,* and *HD-2*) also were identified and analyzed in the assembled genome of *P. pachyrhizi* isolate 71G2. The mating-type genes sequence of 37 other *P. pachyrhizi* isolates from different Brazilian states (Bahia, Mato Grosso, Minas Gerais, Paraná, Rio Grande do Sul, and Santa Catarina) and countries (Paraguay and Japan) were aligned to search for possible alleles. The mating-type loci from *P. meibomiae* and *Phakopsora* sp. were also identified by the tblastn tool implemented locally using the previously *P. pachyrhizi* annotated sequences (Morgulis et al., 2008).

In these studies, the nomenclature followed the names proposed by Cuomo et al., (2017). The protein identification (Protein Id) from the reference genome of isolate PPUFV02 (https://mycocosm.jgi.doe.gov/PpacPPUFV02/PpacPPUFV02.info.html) is listed on the supplementary table 1.

### 2.5 Prediction of domains and three-dimensional structures of mating type proteins

For three-dimensional structures, it was used a combination of MMseqs2 and AlphaFold2 programs, according to the recommendations in their manuals (Jumper et al., 2021; Mirdita et al., 2022). The Interproscan (https://www.ebi.ac.uk/interpro/), SignalP5 (Almagro et al., 2019), and Protter tools (https://wlab.ethz.ch/protter/) allowed the prediction of the protein domains.

### 2.6 Phylogenetic analyses

The amino acids sequences (Supplementary Table S1) were aligned by the MAFFT v.7 (Katoh et al., 2019) with default parameters, and the alignment curation was performed using BMGE (Criscuolo et al., 2010). The phylogenetic relationships were performed for STE3-like proteins and HD1 and HD2 proteins. The STE3-like proteins were C-terminally truncated to optimize the alignment. Phylogenetic inferences were reconstructed using Bayesian inference (BI) analyses, with the evolution models selected using the parameter *prset aamodelpr=mixed* in Mrbayes v. 3.2.7 (Ronquist et al., 2012). Eight runs with four Markov Chain Monte Carlo (MCMC) simulations were conducted simultaneously using 10 million generations, starting from random initial trees. The best amino acid model was selected. The model for each set of data was Cprev for STE and Jones for HD. The HD final tree was rooted at the midpoint and the STE3 tree was rooted at *Ustilago hordei* PRA1 in the iTOL platform (Letunic et al., 2021).

## 3. RESULTS

### 3.1 Genome assembly and mating-type gene identification

The genome of *P. pachyrhizi* isolate 71G2, *Phakopsora* sp. isolate MG19.8.2 and *P. meibomiae* isolate MG19.3.5 were sequenced, assembled, and compared with PPUFV02 genome assembly metrics obtained by Gupta et al. (2022) (Table 1). The genome sequences of the *P. pachyrhizi* isolates showed similar sizes to that of PPUFV02 isolate exhibiting 1,273 Mbp and the 71G2 isolate 1.411 Mbp. For *Phakopsora* sp. and *P. meibomiae* isolates a smaller genome was observed with approximately 790 Mbp in size, in dikaryotic assemblies. About the assembled parameters, the longest contigs and N50 were obtained for MG19.8.2 and MG19.3.5 assemblies (Table 1).

The genomes' assembly completeness analysis performed in BUSCO showed that more than 90% of gene sequences were complete, with the majority of these complete gene sequences categorized as complete and duplicated for all isolates analyzed (Table 1). The completeness of the PPUFV02 assembly also was assessed, resulting in a value of 89,9% of completeness, slightly lower than obtained for the 71G2 isolate (90,4%) (Table 1).

**Table 1:** Genome assembly and BUSCO summary result for *Phakopsora* sp. isolate MG19.8.2, *P. meibomiae* isolate MG19.3.5 and *P. pachyrhizi* isolates 71G2 and PPUFV02.

| **Genome assembly** | | | | |
|---|---|---|---|---|
| | **Isolates** | | | |
| | **71G2** | **MG19.3.5** | **MG19.8.2** | **PPUFV02*** |
| Contig number | 7.420 | 3.758 | 5.987 | 3.140 |
| Max contig length (bp) | 5.612.061 | 15.139.508 | 34.574.969 | 4.158.533 |
| N50 length (bp) | 884.602 | 3.261.550 | 7.661.707 | 677.464 |
| L50 | 410 | 65 | 24 | 556 |
| Total length (Mbp) | 1.411,57 | 789,42 | 791,91 | 1.273,66 |
| GC% | 37,30 | 39,86 | 43,56 | 37,14 |
| **Completeness of the genome assembly (BUSCO)** | | | | |
| | **Isolates** | | | |
| | **71G2** | **MG19.3.5** | **MG19.8.2** | **PPUFV02** |
| Complete total (S + D) | 1.594 (90,4%) | 1.614 (91,5%) | 1.638 (92,9%) | 1.586 (89,9%) |
| Complete and single-copy (S) | 111 (6,3%) | 86 (4,9%) | 176 (10%) | 398 (22,6%) |
| Complete and duplicated (D) | 1.483 (84,1%) | 1.528 (86.6%) | 1.462 (82,9%) | 1.188 (67,3%) |
| Fragmented | 21 (1,2%) | 23 (1,3%) | 19 (1.1%) | 21 (1,2%) |
| Missing | 149 (8,4%) | 127 (7,2%) | 107 (6%) | 157 (8,9%) |
| Total | 1.764 (100%) | 1.764 (100%) | 1.764 (100%) | 1.764 (100%) |

*PPUFV02 genome assembly and metrics obtained by Gupta et al., (2023).

Three *P. pachyrhizi* isolates (PPUFV02, MT2006, and K8108) were recently assembled and characterized by Gupta et al. (2023). The sequences were made available on the Mycocosm website and were here compared with the 71G2 isolate assembly.

Unfortunately, all *P. pachyrhizi* isolates assemblies (K8108, MT2006, PPUFV02, and 71G2) exhibited a high number of contigs that resulted in breaks, fragmenting the assemblies, and sometimes unconnected linked genes. Therefore, the genomes from isolates PPUFV02 and 71G2, which showed the lowest L50 and the highest N50 among the four *P. pachyrhizi* genome assemblies, were selected to identify the mating-type genes and understand how they are linked.

The search for mating-type proteins in *P. pachyrhizi* was performed based on similarity to *P. graminis* f.sp. *tritici* mating-type protein sequences. The two alleles of each *Hd1* and *Hd2* gene, the *Ste3.2-1* gene with its two alleles, the *Ste3.2-2* and *Ste3.2-3* genes with just one allele, and the duplicated *mfa3* gene were identified. Only the *mfa2* gene was not found in *P. pachyrhizi* genome using blast tools. Since *mfa2* usually is linked to the *Ste3.2-2* gene, a manual search was performed for expressed genes that could encode the pheromone precursor in the PPUFV02 genome, using the IGV browser. The putative peptides with C-terminal CAAX domain were selected and aligned against the NCBI proteins database using Blastp. One putative *mfa2* gene approximately 1,2 Mb apart from *Ste3.2-2* at PPUFV02 in contig 6, represented once in the genome was identified. This candidate *mfa2* also showed 54,1% similarity with a hypothetical protein from *Austropuccinia psidii* (MBW0494438.1) that contains pheromone precursor features and 32,91% similarity with a *M. larici-populina* putative pheromone peptide (XP_007415087.1). After this identification process, all mating-type genes were manually annotated using the predicted proteins and the cDNA reads from the RNA-seq data available at the Mycocosm platform to support exon regions (Supplementary file 1).

For, MG19.8.2 and MG19.3.5 isolates, the mating-type genes were identified using the previously annotated genes for *P. pachyrhizi* using the tblastn tool. Due to the genetic proximity among the *Phakopsora* species, all *Ste3.2* and *HD* genes were identified by the tblastn tool. After the identification of the genes, all mating-type genes in both *Phakopsora* sp. and *P. meibomiae* isolates were manually annotated. The same number of gene copies and alleles for *Ste3.2* and *Hd* genes were observed in all *Phakopsora* species (*P. pachyrhizi, P. meibomiae,* and *Phakopsora* sp). However, for *mfa2* and *mfa3* genes, variations in the number of copies and alleles were identified in *Phakopsora* sp. and *P. meibomiae* species when compared with *P. pachyrhizi* isolates. While *P. pachyrhizi* harbors one copy of *mfa2* and two of *mfa3*, *P. meibomiae* has six paralogs of the *mfa2* and four of *mfa3,* and in *Phakopsora* sp. isolate was observed four paralogs of *mfa2* and three of *mfa3* were observed.

### 3.2 Phylogenetic Analyses

Molecular phylogenetic relationships were performed separately for STE3.2 and HD proteins, from ten species of *Pucciniales* order, including *Cronartium quercuum* f. sp. *Fusiforme*, *Melampsora larici-populina*, *Puccinia graminis* f. sp. *Tritici*, *P. striiformis* f. sp. *Tritici*, *P. triticina* 1-1 BBBD Race 1, *Phakopsora pachyrhizi*, *Austropuccinia psidii*, *Sphaerophragmium acaciae*, *Phakopsora* sp. and *P. meibomiae* (Supplementary table S1) (Figure 2). *Ustilago hordei* PRA1 (STE3.1) sequence was used as the outgroup in STE3.2 analysis and no outgroup was used in HD analysis since both proteins (HD1 and HD2) were used in phylogenetic reconstruction. The STE3 tree grouped the protein sequences in two big clades: the first clade included exclusively the biallelic STE3.2-1 (represented by one allele) and the second clade with the others STE3.2 sequences that were grouped in subclades for STE3.2-2, STE3.2-3, and STE3.2-4 (Figure 2A; blue boxes). The STE3.2-1 protein is present in all ten analyzed rust species. The STE3.2-3 also is present in all species, except for *C. quercuum* f. sp. *fusiforme*. The STE3.2-2, on the other hand, seems to be absent in *M. larici-populina* and *C. quercuum* f. sp. *fusiforme,* species that contain STE3.2-2 and STE3.2-4 proteins grouped in the subclade STE3.2-4 that contains two similar protein sequences of each species.

The phylogenetic HD tree was rooted at the midpoint, supporting two distinct clades that corresponded to each HD protein (Figure 2B). For each HD protein, four subclades were observed that included sequences of species from the same family or related families: a) Pucciniaceae; b) Phakopsoraceae; c) Sphaerophragmiaceae; and d) Melampsorineae and Cronartiaceae (*M. larici-populina* and *C. quercuum* f. sp. *fusiforme*, respectively).

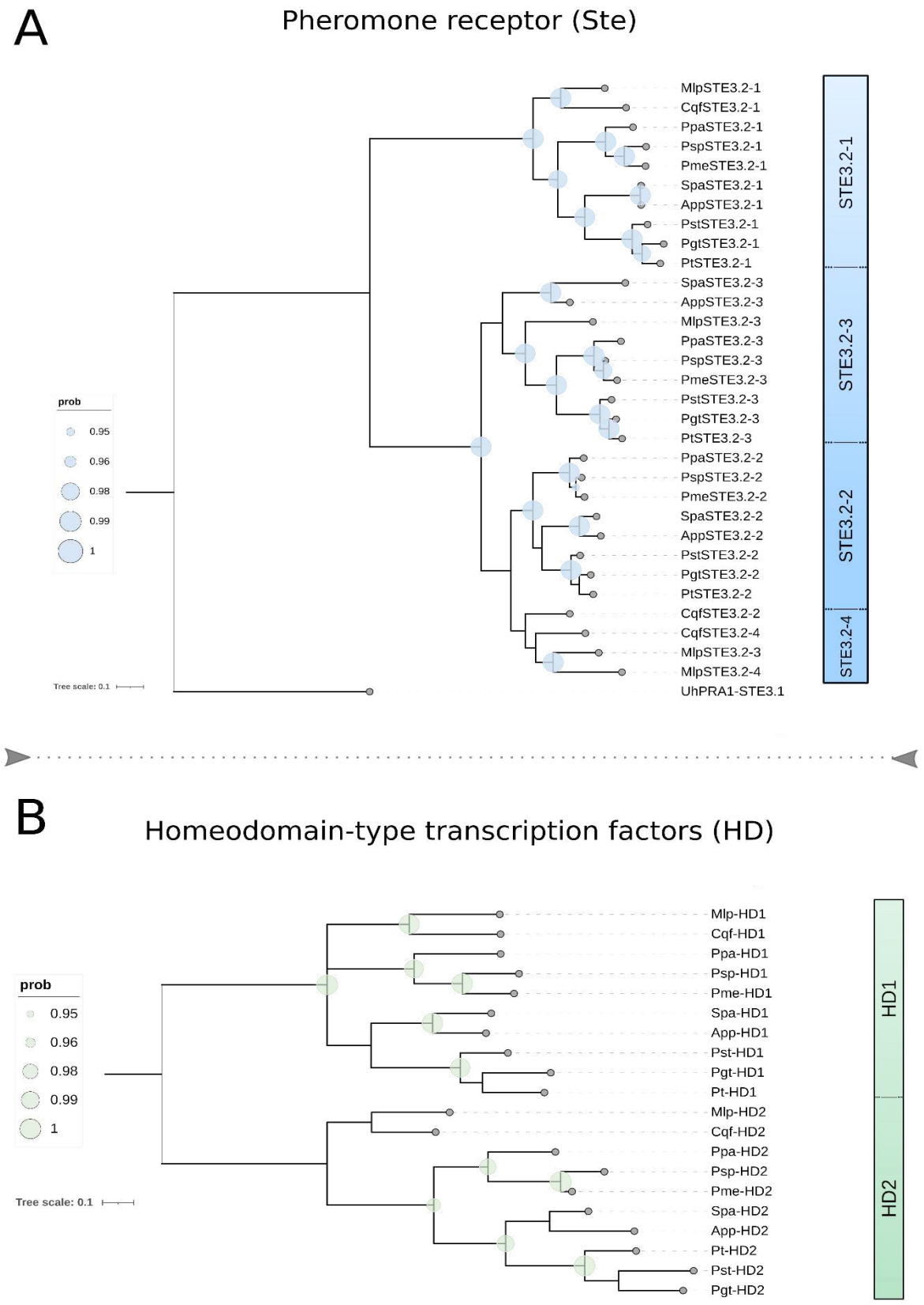



**Figure 2**: Phylogenetic relationship of STE3-like pheromone receptor proteins and mating-type homeodomain-containing transcription factors from ten rust fungi (*Pucciniales*). A) Molecular phylogenetic relationship of STE3-like pheromone receptor proteins (STE3.2-1, STE3.2-2, STE3.2-3,

and STE3.2-4) from Pucciniales species that hold a distinct number of STE3.2 homologs and *Ustilago hordei* STE3.1, used as outgroup. The pheromone receptor sequences were C-terminally truncated to optimize the alignment. B) Midpoint rooted tree displaying the molecular phylogenetic relationship of mating-type homeodomain transcription factors HD1 and HD2. Cqf: *C. quercuum* f. sp. *fusiforme*; Mlp: *M. larici-populina*; Ppa: *P. pachyrhizi*; Psp: *Phakopsora* sp.; Pme: *P. meibomiae*; Pgt: *P. graminis* f. sp. *tritici*; Pst: *P. striiformis* f. sp. *tritici*; Pt: *P. triticina*; App: *Austropuccinia psidii*; Spa: *Sphaerophragmium acaciae*; Uh: *U. hordei*. Phylogenetic analyses were performed using the Bayesian inference (BI) analyses, based on the Markov Chain Monte Carlo (mcmc) method; circles indicate the posterior probabilities and branch lengths are proportional to the number of substitutions per site. Additional information is presented in Supplementary Table 1.

## 3.3 Analysis of *Phakopsora* species mating-type system

### 3.3.1 Disposition of mating-type genes on *Phakopsora* spp. genomes

*P. pachyrhizi* has the largest and most repetitive genome among the analyzed species (Table 1) and these features resulted in the less contiguous assembly among *Phakopsora* species. The assembly of 71G2 and PPUFV02 genomes, the most contiguous assemblies of *P. pachyrhizi*, were used to characterize the mating-type structure of the species. The analysis and comparison of both genomes of isolates PPUFV02 and 71G2 made it possible to identify more physically linked mating-type genes. Later, we also confirmed the presence of the same mating-type structure and identical sequences in both isolates K8108 and MT2006 but distributed among more contigs.

The *HD* locus in *P. pachyrhizi*, harboring *Hd1* and *Hd2*, was identified in a region of approximately 3 Kb in two different contigs (Figure 3). Each contig harbors one specific allele of each gene that is identical (nucleotide sequence) among the other 37 *P. pachyrhizi* isolates analyzed (data not shown). Genes *Hd1* and *Hd2* are transcribed in opposite directions and RNA-seq transcripts data support the expression of both HD genes. In both MG19.8.2 (*P. meibomiae*) and MG19.3.5 (*Phakopsora* sp.), the *HD* loci were also identified in a region of approximately 3 Kb, in two different contigs, containing one specific allele of each gene (Figure 3). Genes *Hd1* and *Hd2* of *P. meibomiae* and *Phakopsora* sp. also are transcribed in opposite directions and both contigs containing *Hd* loci are not linked to any other contigs containing mating-type genes in these genome assemblies (Figure 3).

The gene *Ste3.2-1* is present in two different contigs among all *Phakopsora* species analyzed, compounding two possible alleles, since their encoded protein's similarity is higher than 98%. The gene *Ste3.2-1* was not physically linked to any other mating-type gene (Figure 3). The genes *Ste3.2-2* and *Ste3.2-3* were identified once in each assembly and they were linked to *mfa2* and *mfa3* genes, respectively, but the gene orientation, distance, and the

number of paralogs vary between *Phakopsora* species, demonstrating variation in the mating-type structure (Figure 3).

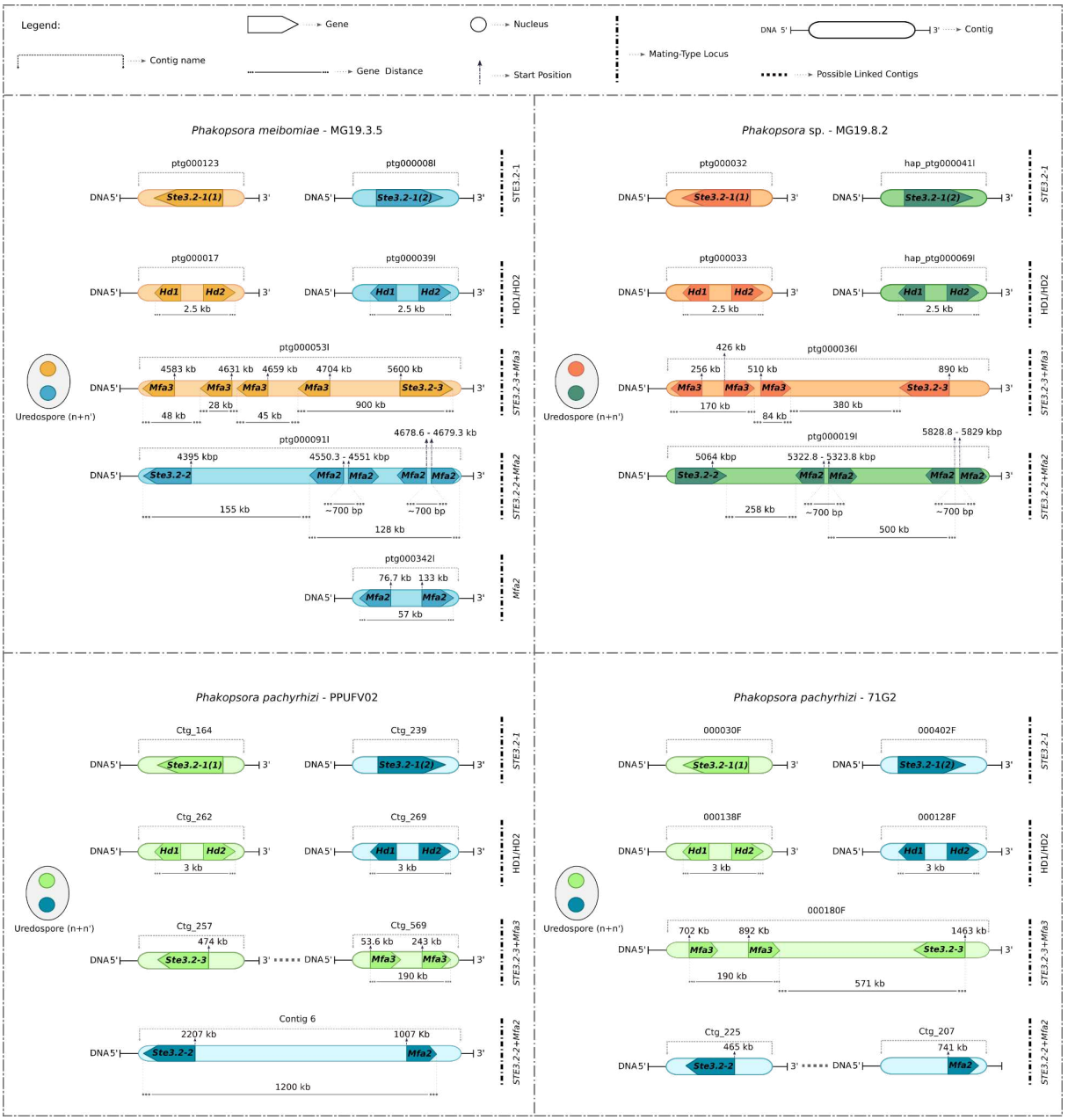



**Figure 3:** Genomic organization of mating type loci from the dikaryotic fungi *P. pachyrhizi*, *Phakopsora* sp., and *P. meibomiae* inferred from genome assemblies. Different colors were used to distinguish that each mating set is associated with one nucleus in the dikaryotic cell. Contigs marked with the same color could be associated with the same nucleus. Equivalent contigs of *P. pachyrhizi* isolates PPUFV02 and 71G2 also were marked with the same color. The *HD* loci and *STE3.2-1* genes are biallelic and they are present in both nuclei, while locus *Ste3.2-2/mfa2* and *Ste3.2-3/mfa3* are supposed to be nucleus-specific.

In *P. pachyrhizi* PPUFV02 assembly, *Ste3.2-2* is located in the Contig 6, one of the largest contigs of the assembly (~3,56 Mbp), that also harbors the *mfa2* gene approximately 1,2 Mb apart. In all other *P. pachyrhizi* genome assemblies, these genes are located in smaller contigs (ex. Contigs 207 and 225 at the genome assembly of isolate 71G2) where the connection between *Ste3.2-2* and *mfa2* genes could not be observed (Figure 3). A similar situation occurred with *Ste3.2-3* and the closest copy of *Mfa3* genes that are linked, approximately 580 Kb apart, but this connection is only observed in the assembly of isolate 71G2 where those genes are located in the larger Contig 180F, while in PPUFV02 these genes are separated in two small contigs (contigs 257 and 569) (Figure 3). The *mfa3* gene in contigs of PPUFV02 and 71G2 assemblies is located in one duplicated region of the genome that resulted in two identical copies of *mfa3* separated for 190 Kb. Putting those observations together, it is possible to infer that both pheromone receptor genes are linked to one pheromone precursor gene. All *Ste3.2* and *mfa* genes had their expression confirmed by analysis of RNA-seq data.

The smaller genome of *P. meibomiae* and *Phakopsora* sp. compared with *P. pachyrhizi* enabled better genome assemblies (Table 1). *Phakopsora* sp. and *P. meibomiae* assemblies also have the pheromone precursor (*mfa*) and pheromone receptor genes (*Ste3.2*) in the same contig, but these species have multiple paralogs of the *mfa2* and *mfa3* genes, in variable arranges and copy number (Figure 3). In *P. meibomiae*, four paralogs of the *mfa3* gene were found in a region of approximately 120 Kb between the first and the last paralog. All *mfa3* are in the same orientation and the closest copy to the *Ste3.2-3* gene is around 900 Kb apart. In *Phakopsora* sp., the three paralogs of the *mfa3* gene are in a region of approximately 255 Kb, with the genes oriented in different directions, and 380 Kb apart from the gene *Ste3.2-3* (Figure 3).

*Mfa2* has a different disposition from the *mfa3* gene. In both MG19.8.2 (*Phakopsora* sp.) and MG19.3.5 (*P. meibomiae*) isolates, the *mfa2* genes, linked to *Ste3.2-2*, have four paralogs separated in two pairs. The copies of each pair are oriented in opposite directions and separated for approximately 0,7 Kb. In *Phakopsora* sp., these two pairs of *mfa2* paralog genes are separated for 500 Kb and the closest copy of the *Ste3.2-2* is around 258 Kb of distance, while in *P. meibomiae*, the pairs of *mfa2* genes are 128 Kb apart and the closest copy is around 155 Kb from the gene *Ste3.2-2*. Additionally, two paralogs of the *mfa2* gene are located in another contig of *Phakopsora* sp., maintaining opposite orientations, but 57 Kb apart (Figure 3).

### 3.3.2 Mating-type protein analyses

The pheromone receptors STE3.2 are typical seven transmembrane receptors (7TM) coupled to G-proteins. The Protter, Interproscan, and Alpha Fold programs were used to confirm the presence of the domains and structure for these proteins in *Phakopsora* species and *P. triticina*. Interproscan was used to identify conserved domains in STE3.2 proteins. The seven typical transmembrane domains were identified at most of the STE3.2-2 and STE3.2-3 receptors, but not in STE3.2-2 of *P. pachyrhizi*, where just five of the transmembrane domains were identified (Supplementary Figure 1). Interproscan also identified one possible signal peptide sequence in the STE3.2-2 predicted protein of *P. pachyrhizi*. Protter, which uses Phobius to identify proteins' domains, also reported an atypical structure for STE3.2-2 of *P. pachyrhizi* with five transmembrane domains and a signal peptide domain (Figure 4). Aiming to check the presence of signal peptide in *STE3.2-2* encoded protein of *P. pachyrhizi*, the SignalP 5.0 program was used, but no secretion signal was predicted (Supplementary Figure 1).

The predicted structure of proteins STE3.2 from *Phakopsora* species indicates that they are more compact than the proteins STE3.2 of *P. triticina* (Figure 4). The predicted proteins STE3.2-2 and STE3.2-3 of *Phakopsora* species were very similar, with differences in the number of transmembrane helices and in the length of intracellular C-terminal tail (Figure 4). These transmembrane domains demonstrated a clear membrane association of these receptors' proteins (Figure 4). All analyzed proteins, including *P. triticina* ones, showed six loop regions separated by parallel regions with alpha-helix conformation where the transmembrane domains predicted by Interproscan and Protter are located. Although STE3.2-2 of *P. pachyrhizi* also has these seven alpha-helix regions, only five of them were predicted to have transmembrane domains. The number of beta-sheet and alpha-helix regions is similar in STE3.2 proteins, with small differences in the secondary conformation between the 7TM receptors STE3.2-2 and STE3.2-3 concentrated at the N-terminal region, where the STE3.2-3 proteins and also the STE3.2-2 of *P. triticina* do not have the alpha-helix secondary structure that is present in the predicted STE3.2-2 proteins of the *Phakopsora* spp. (Figure 4).

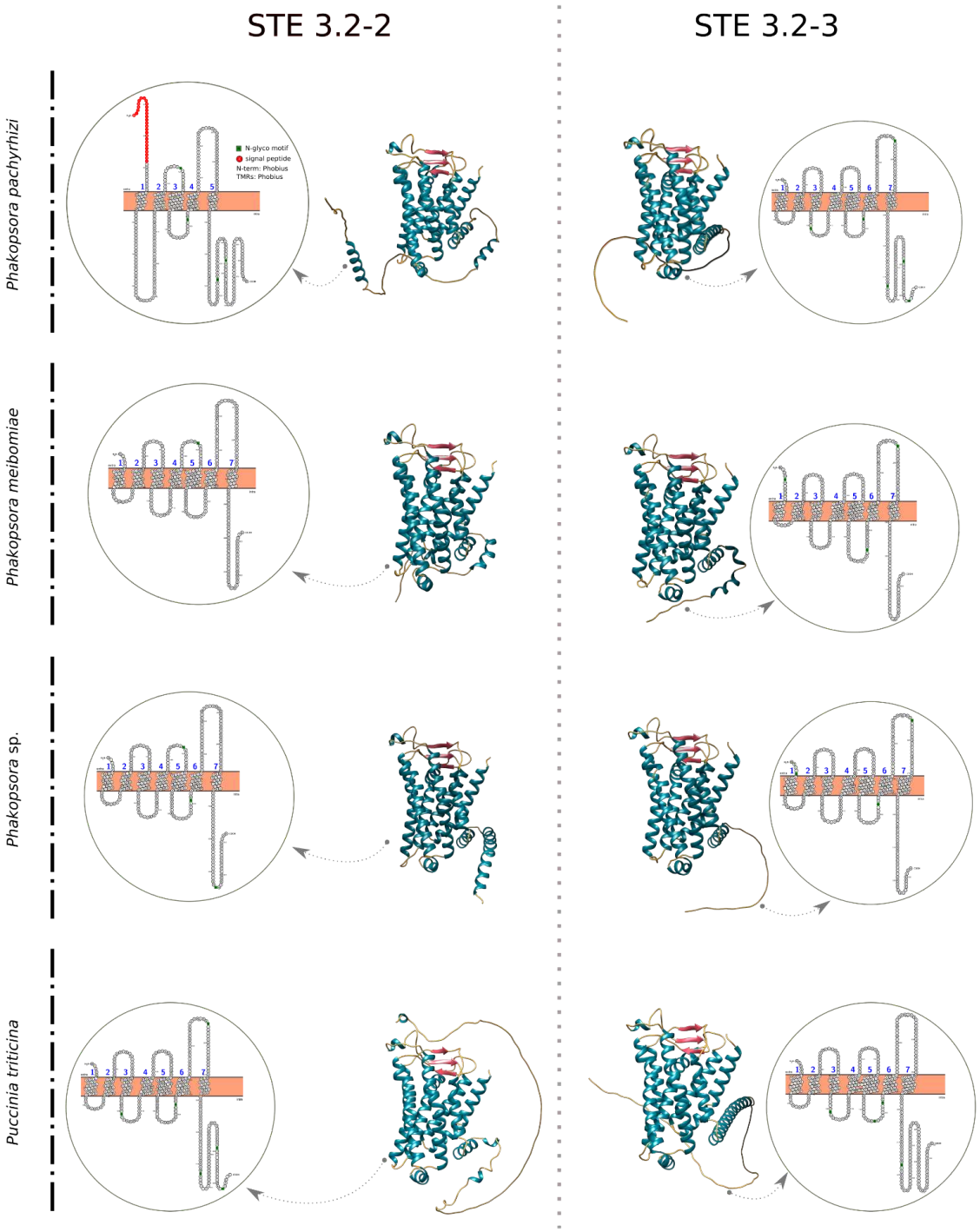



**Figure 4**: Tertiary and secondary predicted structure of the seven transmembranes (7TM) pheromone receptor proteins STE3.2-2 and STE3.2-3 from *Phakopsora* species and *P. triticina*. The secondary predictions were performed using amino acid sequences in Protter and the tertiary structures were predicted in Alpha Fold. Atypical 7TM pheromone receptor protein with five transmembrane domains was predicted for STE3.2-2 of *P. pachyrhizi*, while the typical 7TM was predicted for the other proteins. In blue were presented the alpha-helices and in red the Beta-sheets.

The predicted pheromone precursors Mfa2 and Mfa3 of *P. pachyrhizi* have 87 and 80 amino acids in length, respectively (Figure 5A). Mfa2 is encoded by a single copy gene and Mfa3 could be encoded by two identical copies of the *mfa3* (Figure 3; Figure 5A, orange lines). Mfa2 and Mfa3 share 39% of similarity (Figure 5B, orange vertices). The homologs of these peptides in *Phakopsora* sp. and *P. meibomiae* have between 84 and 89 amino acids. All predicted Mfa2 and Mfa3 peptides exhibited features related to pheromone precursors, like tandem repeats (KELGGSNH) and CAAX-domains in their N-terminal region.

*Phakopsora* sp. has two pairs of *mfa2* genes. Each pair encodes one variant of the Mfa2 (Figure2; Figure 5A, green line), that shares 82,14% similarity between them. *P. meibomiae* also has two pairs of *mfa2* genes, but in this species, there is just one variant with a substitution of one amino acid (G60D) (Figure 5A, blue line). Additionally, *P. meibomiae* has two other *mfa2* genes located in another contig. One of these copies has a premature stop codon after the 22nd amino acid, resulting in a truncated peptide (Figure 5A, blue line). The second additional *mfa2* is not truncated and possibly encodes a functional peptide, that exhibited 94% similarity with the other paralogs. *P. meibomiae* has four identical copies of the predicted peptide Mfa3 (Figure 5A, blue line), while *Phakopsora* sp. has three different paralogs with divergent amino acids in two positions (A49S and Y74H) (Figure 5A, green line).

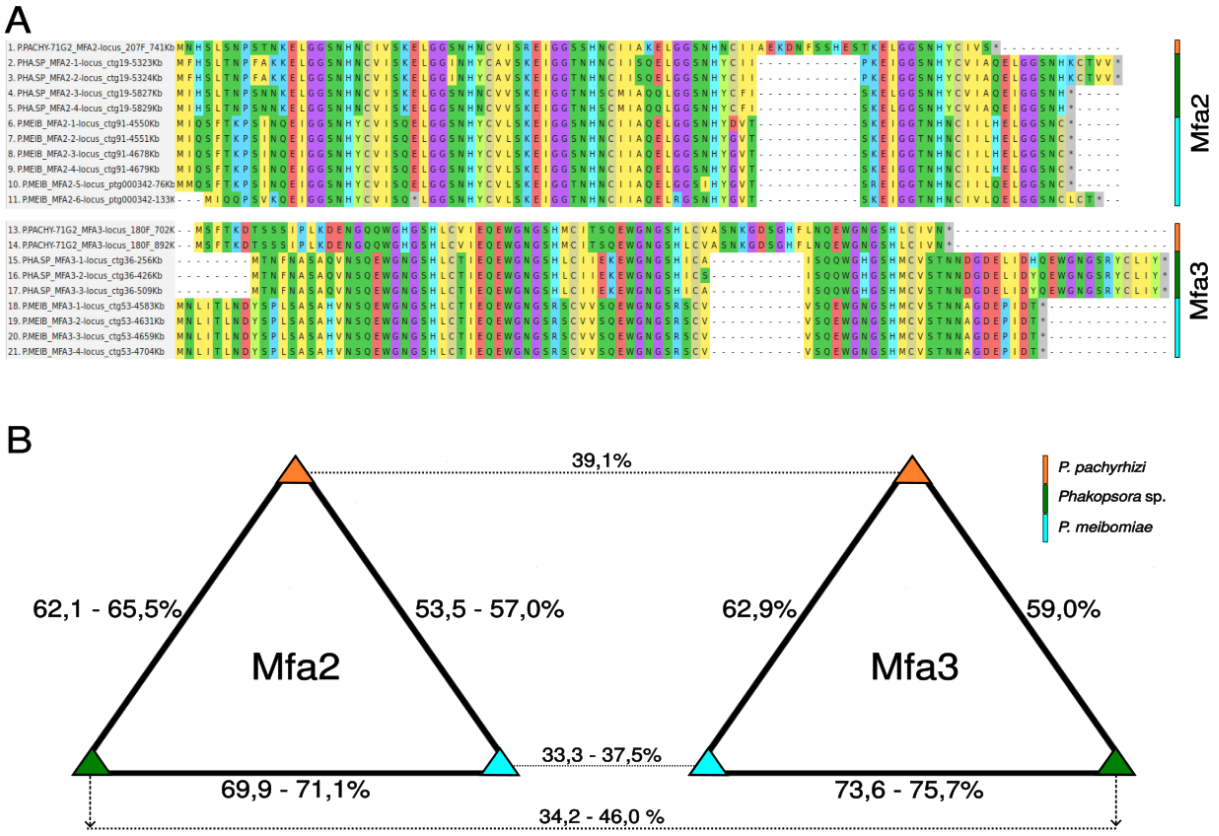



**Figure 5:** Sequence comparison and analysis of similarities among Mfa2 and Mfa3 proteins from *Phakopsora* species. A) Amino acid sequences alignment of Mfa2 proteins (lines 1-11) and Mfa3 proteins (lines 13-21) from different *Phakopsora* species were aligned separately to highlight their inter and intra-specific differences; B) Values or intervals of similarity in pairwise comparations between Mfa2 and Mfa3 peptides were represented in two triangles. The colored vertices of triangles represent *Phakopsora* species and lines connect the species being compared. Continuous lines link comparisons from the same Mfa peptide in different species and dashed lines link comparisons between different Mfa peptides in the same species.

The similarity comparisons between the predicted peptides Mfa2 and Mfa3 in the same species exhibited values lower than 50% (Figure 5B). Mfa2 from *Phakopsora* sp. is around 70% similar to Mfa2 of *P. meibomiae*, while the peptides Mfa3 share about 73% similarity. When compared with *P. pachyrhizi*, Mfa2 peptides from *P. meibomiae* and *Phakopsora* sp. have, respectively, about 53% and 62% of similarity, and for Mfa3 peptide these values were slightly bigger: approximately 63% and 59% of similarity (Figure 5B).

## 4. DISCUSSION

Some rust fungi can reproduce sexually and asexually, using one or more hosts. Both *P. pachyrhizi* and *P. meibomiae* infect more than fifty hosts (Ono et al., 1992; Slaminko et al., 2008), but no sexual reproduction has been reported in these species yet. In the case of *P. pachyrhizi*, studies indicate that populations of South America have been reproducing exclusively asexually (Jorge et al., 2014; Darben et al., 2020), a fact that is corroborated by the high heterozygosity between the nuclei observed in clonal species (Gupta et al., 2023). Here, the mating-type system of both species and one possible new species of *Phakopsora* sp. that infects soybean were analyzed, as a way of clarifying the structure and if the system is still functional even with the absence of sexual reproduction.

Mating-type genes were identified in the genome assembly of four isolates of *P. pachyrhizi* (PPUFV02, MG2006, K8101, and 71G2), one isolate of *P. meibomiae* (MG19.3.5) and one isolate of *Phakopsora* sp. (MG19.8.2). The mating-type system in *Phakopsora* contains one contig harboring *Ste3.2-2*/*mfa2* and a second contig harboring *Ste3.2-3*/*mfa3*. The *pheromone receptor gene* (*Ste3.2*) is present once per contig, while the *pheromone precursor genes* (*mfa*), with the exception of *P. pachyrhizi mfa2*, are present more than one time in the same contig. Additionally, *P. meibomiae* showed paralogs of *mfa2* located in two contigs that do not seem to be linked. The *HD locus* exhibited both biallelic *homeodomain-containing transcription factor genes,* which were not linked to *Ste/mfa* loci. The biallelic gene *Ste3.2-1* was unlinked to any other mating-type gene, it showed a low level of intraspecific polymorphism and probably it was unrelated to mating compatibility (Coelho et al., 2017). Small-reads DNA sequence data from 37 *P. pachyrhizi* isolates were aligned, including samples from Brazil, Paraguay, and Japan, but no variation was observed in the structure of the mating type loci in these isolates when compared with other isolates previously cited, and just two mating variants were identified in our samples (unpublished data). It indicates a conserved system and a population structured with just two mating variants, possibly due to limited migration events of *P. pachyrhizi* from its center of origin in Asia to South America. However, further studies with a higher number of samples and with the population from the center of origin of the species have to be done to confirm if there are more mating variants than these two detected in our samples. *Phakopsora* spp. mating-type system organization is similar to the system present in *Puccinia triticina*, *P. striiformis* f.sp. *tritici* and *P. graminis* f.sp. *tritici* with: a) two biallelic *Hd* genes unlinked to *Ste* or *mfa* genes; b) one biallelic gene *Ste3.2-1*, not linked to any mating-type gene, and probably not a mating

determinant; and c) one copy of the mating-specific gene *Ste3.2-2* and *Ste3.2-3*, linked to putative *mfa* genes, also mating-specific (Cuomo et al., 2017). This *MAT* loci organization supports one possible tetrapolar system where two variant loci (*Hd* and *Ste3/mfa*) would be unlinked and their segregation would occur independently (Figure 3).

Although the tetrapolar system seems to be conserved, remarkable differences are present in pheromone precursor and pheromone receptor genes. Each putative *mfa(2/3)* gene is linked to one *Ste3.2(2/3)* gene, as observed in *Puccinia* spp., but the distance between these genes is significantly larger in *Phakopsora* spp. than in *Puccinia* spp. The *mfa2* gene identified by Cuomo et al. (2017) was between 0,5 and 0,7 Kb apart from *Ste3.2-2* in all analyzed *Puccinia* spp., and one putative *mfa* gene of *P. graminis* f.sp. *tritici* was 24 Kb away from *Ste3.2-3*. In *P. pachyrhizi*, the *Ste3.2-2* and *Ste3.2-3* genes are approximately 1,2 Mb and 580 Kb respectively, from the putative *mfa2* and *mfa3* (Figure 3). The larger distance between *mfa* and *Ste3.2* genes also was observed in *Phakopsora* sp. and *P. meibomiae.* These variations could be related to the expansion of the genome size and the presence of repetitive elements reported by Grupta et al. (2023) for *P. pachyrhizi*, features which also were present among other *Phakopsora* spp. analyzed. This genome expansion in repetitive content included high amounts of transposable elements could have contributed to genome rearranges that cause gene duplications as observed for some *mfa* genes. The increase in distance between two loci also could in the future displace previously closed genes in the genomes and enable more recombination events between them, creating new mating combinations.

Our phylogenetic analysis also supports the proximity between *Phakopsora* and *Puccinia* STE3.2 and HD proteins (Figure 2). *Puccinia* spp. showed the same number of genes and the most similar proteins with *Phakopsora* species. However, the amino acid substitutions in *Phakopsora* spp. concerning *Puccinia* presented a significant impact on STE3.2 protein structures, suggesting probable alterations in recognition and signaling domains. While all predicted STE3.2 proteins of *Puccinia* spp. reported by Cuomo et al. (2017) exhibited the typical seven transmembrane-receptor protein conformation with extracellular N-terminal portion, C-terminal intracellular portion, and seven alpha-helices transmembrane domains (identified using Protter), *P. pachyrhizi* STE3.2-2 could have an atypical structure conformation, with only five transmembrane domains (Figure 4). Protein domains predicted with Interproscan and Protter indicated a possible loss of transmembrane domains and a consequent change in protein conformation. Protein structure predicted with AlphaFold indicates that *Phakopsora* sp. and *P. meibomiae* predicted proteins STE3.2-2 and STE3.2-3 have the conserved structure of the 7TM, but with more compact structure (closest

alpha-helices) and with modification on the secondary conformation on the N-terminal portion of STE3.2-2 protein when compared with the correspondending predicted protein of *P. triticina* (Figure 4).

The predicted putative pheromone precursors (Mfa2 and Mfa3) from *Phakopsora* spp. showed a size between 80 and 87 amino acids in length, demonstrating to be larger than the putative precursors identified in wheat rust pathogens *Puccinia* spp. (with 33 amino acids) by Cuomo et al. (2017), but similar to *M. larici-populina* which predicted Mfa peptides that ranged between 55 and 86 amino acids in size. The larger sequences of *Phakopsora* spp. Mfa2 and Mfa3 are the consequence of a higher number of tandem repeats that were present three or four times in the precursor-predicted sequence, a common characteristic in some annotated pheromone precursors predicted in *M. larici-populina* (Duplessis et al., 2011). Although the proximity among the *Phakopsora* spp. infecting soybeans, for being different species different pheromone precursors were identified, as expected (Figure 5). Variations in peptide sequences, genes' organization, and copy numbers were observed. Surprisingly, *Phakopsora* sp. and *P. meibomiae* exhibited three or more copies of each *mfa* gene, while *P. pachyrhizi* just had one additional copy for *mfa3*. The duplicated region of the *mfa3* gene in *P. pachyrhizi* reveals and conserved peptide without mutations, while *Phakopsora* sp. and *P. meibomiae mfa* genes showed more than one variant, except for the *mfa3* of *P. meibomiae*. This variation highlights the divergent ways of evolution on *mfa* genes of *Phakopsora* species. The duplication of genes usually is associated with gene evolution, leading one of the copies to accumulate mutations while the other one maintains the original function (Zhang, 2003). *P. pachyrhizi* isolates have not accumulated mutations yet, a fact that could be associated with recent duplication of the gene or high fitness and consequent selection of isolates that maintain both copies. On the other hand, changes in amino acid sequences observed among *mfa* genes of the other species could indicate the evolution and accumulation of mutations in some copies while other copies maintain the original function. The increased number of copies could increase the number of peptides produced and the higher number of tandem repeats could be associated with more pheromone molecules processed per unit of peptide, both features conferring a more efficient communication (Coelho et al., 2017). However, distinct paralogs of *mfa* genes encode distinct pheromone peptides that could be or not be recognized by pheromone receptors.

Changes in protein structure and peptide sequence could have strong consequences on protein stability and interaction with other molecules (Prabantu et al., 2021). A single amino acid change in both pheromone precursor peptide and pheromone receptor protein has already

been reported to impair the mating system. Two amino acids deletion at the third transmembrane domain of the pheromone receptor at the *Bβ2* mating-type locus in the basidiomycete *Schizophyllum commune* altered the pheromone receptor recognition and a single amino acid change in the pheromone peptide encoded by the gene *bbp2(1)* also was sufficient to change the mating phenotype (Fowler et al., 2001). Furthermore, several mutations on 7TM receptors in different regions of the proteins, causing amino acid changes or small deletions, were associated with phenotype disorders, as an example of many human diseases (Schöneberg et al., 2004).

In *P. triticina*, *P. graminis f. sp. tritici*, and *P. striiformis f.* sp *tritici* the receptor genes from the Ste3 family were classified as *Ste3.2-1*, *Ste3.2-2*, and *Ste3.2-3*. Usually, *Ste3.2-1* is considered a non-mating-type receptor gene in these species, while *Ste3.2-2* and *3* are likely mating-type pheromone receptor genes and hence nucleus/haplophase-specific (Cuomo et al., 2017). If the protein STE3.2-2 of *P. pachyrhizi* loses its pheromone receptor function by the variations observed, it could be a factor to prevent the mating impairing and consequently, the sexual reproduction, since two haploid cells must carry complementary pheromones and receptors in order to initiate syngamy. In *P. pachyrhizi*, the *Ste3.2-1* also has been identified in a different genomic locus from the other *Ste3.2* genes and it is supposed to not be involved in mating specificity, as supposed for other rust species (Cuomo et al., 2017; Coelho et al., 2017).

The assembled sequences were confirmed in four independent genome assemblies and the RNA-seq data was used to predict the protein. These data support that the protein predictions were correct, but the prediction of domains in some cases were uncertain even between distinct predictors. The amino acid variation on pheromone precursors observed on *Phakopsora* sp. and *P. meibomiae* also could have a significant impact on mating recognition. Single amino acid changes have been reported as sufficient for loss of recognition by pheromone receptors, and pheromone receptors recognizing more than one distinct pheromone also had been reported, indicating that variations on pheromone peptide could have distinct consequences (Fowler et al., 2001). So, to completely understand the real impact of these possible changes in the STE3.2-2 receptor of *P. pachyrhizi* and also the diverse pheromone peptides encoded by *Phakopsora* sp. and *P. meibomiae*, functional analyses must be done to assess and to confirm the interaction between the predicted STE3.2 receptors and pheromone molecules generated from Mfa peptides.

The mating-type system of *Puccinia* spp. studied and predicted by Cuomo et al. (2017) was confirmed four years later when genome assemblies at the chromosome level and

in a phase were obtained (Wu et al., 2021). Unfortunately, the high complexity of the *P. pachyrhizi* genome due to its size (> 1,2Gbp) and repetitive content (>90%) (Grupta et al., 2023), makes it difficult to assemble the genome. Nowadays, no assembly of *Phakopsora* spp. genomes, obtained from dikaryotic uredospores, could separate the DNA that comes from each nucleus and re-assemble the sequenced DNA at the pseudo-chromosome level ("in phase assembly"). For now, it was possible to predict the mating-system organization in *P. pachyrhizi* as tetrapolar, associating the data of this study with the mating-type structure observed in other rust species, and in the future, when higher-quality genome assemblies become available, it will be possible to confirm all these suppositions.

## 5. CONCLUSION

The genomic analysis of *Phakopsora* sp., *P. meibomiae*, and *P. pachyrhizi* mating-type locus supported a tetrapolar mating system with two biallelic *homeodomain-transcription factor genes* and two mating-specific *Ste3/mfa* loci harboring *pheromone receptors* and *pheromone precursor* genes. The system is similar to the reported mating system of *Puccinia* spp.. Atypical variations in the predicted pheromone receptor of *P. pachyrhizi* could lead to loss of function and be a limiting factor for sexual reproduction, but functional studies must be done to confirm the real impact of the protein variation on mating recognition.

## 6. REFERENCES

Almagro Armenteros, J. J., Tsirigos, K. D., Sønderby, C. K., Petersen, T. N., Winther, O., Brunak, S., ... & Nielsen, H. (2019). SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nature Biotechnology*, 37 (4), 420-423.

Bakkeren, G., Kämper, J., & Schirawski, J. (2008). Sex in smut fungi: structure, function and evolution of mating-type complexes. *Fungal Genetics and Biology*, 45, 15-21.

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., & Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature methods*, 18 (2), 170-175.

Coelho, M. A., Bakkeren, G., Sun, S., Hood, M. E., & Giraud, T. (2017). Fungal sex: the Basidiomycota. *Microbiology spectrum*, 5 (3), 5-3.

Criscuolo, A., & Gribaldo, S. (2010). BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC evolutionary biology*, 10, 1-21.

Cuomo, C. A., Bakkeren, G., Khalil, H. B., Panwar, V., Joly, D., Linning, R., ... & Fellers, J. P. (2017). Comparative analysis highlights variable genome content of wheat rusts and divergence of the mating loci. *G3: Genes, Genomes, Genetics*, 7 (2), 361-376.

Darben, L. M., Yokoyama, A., Castanho, F. M., Lopes-Caitar, V. S., da Cruz Gallo de Carvalho, M. C., Godoy, C. V., ... & Marcelino-Guimarães, F. C. (2020). Characterization of genetic diversity and pathogenicity of *Phakopsora pachyrhizi* mono-uredinial isolates collected in Brazil. *European Journal of Plant Pathology*, 156, 355-372.

Duplessis, S., Cuomo, C. A., Lin, Y. C., Aerts, A., Tisserant, E., Veneault-Fourrey, C., ... & Martin, F. (2011). Obligate biotrophy features unraveled by the genomic analysis of rust fungi. *Proceedings of the National Academy of Sciences*, 108 (22), 9166-9171.

Duplessis, S., Bakkeren, G., & Hamelin, R. (2014). Advancing knowledge on biology of rust fungi through genomics. *In Advances in botanical research*, 70, 173-209.

Duplessis, S., Lorrain, C., Petre, B., Figueroa, M., Dodds, P. N., & Aime, M. C. (2021). Host adaptation and virulence in heteroecious rust fungi. *Annual Review of Phytopathology*, 59, 403-422.

Esser, K. (1971). Breeding systems in fungi and their significance for genetic recombination. *Molecular and General Genetics*, 110 (1), 86-100.

Ferrarezi, J. A., McTaggart, A. R., Tobias, P. A., Hayashibara, C. A., Degnan, R. M., Shuey, L. S., ... & Quecine, M. C. (2022). *Austropuccinia psidii* uses tetrapolar mating and

produces meiotic spores in older infections on *Eucalyptus grandis*. *Fungal Genetics and Biology*, 160, 103692.

Foulongne-Oriol, M. (2012). Genetic linkage mapping in fungi: current state, applications, and future trends. *Applied Microbiology and Biotechnology,* 95 (4), 891-904.

Fowler, T. J., Mitton, M. F., Vaillancourt, L. J., & Raper, C. A. (2001). Changes in mate recognition through alterations of pheromones and receptors in the multisexual mushroom fungus *Schizophyllum commune*. *Genetics*, 158 (4), 1491-1503.

Goellner, K., Loehrer, M., Langenbach, C., Conrath, U. W. E., Koch, E., & Schaffrath, U. (2010). *Phakopsora pachyrhizi*, the causal agent of Asian soybean rust. *Molecular plant pathology*, 11 (2), 169-177.

Gupta, Y. K., Marcelino-Guimarães, F. C., Lorrain, C., Farmer, A. D., Haridas, S., Ferreira, E. G. C., ... & van Esse, H. P. (2023). Major proliferation of transposable elements shaped the genome of the soybean rust pathogen *Phakopsora pachyrhizi*. *Nature Communications*, 14, 1835.

Hartman, G. L., Rupe, J. C., Sikora, E. J., Domier, L. L., Davis, J. A., & Steffey, K. L. (Eds.). (2015). *Compendium of soybean diseases and pests*. St. Paul, MN: American Phytopathological Society.

Heitman, J., Sun, S., & James, T. Y. (2013). Evolution of fungal sexual reproduction. *Mycologia*, 105 (1), 1-27.

Hibbett, D. S., Binder, M., Bischoff, J. F., Blackwell, M., Cannon, P. F., Eriksson, O. E., ... & Zhang, N. (2007). A higher-level phylogenetic classification of the Fungi. *Mycological research*, 111 (5), 509-547.

Jorge, V. R., Silva, M. R., Guillin, E. A., Freire, M. C. M., Schuster, I., Almeida, A. M. R., & Oliveira, L. O. (2015). The origin and genetic diversity of the causal agent of Asian soybean rust, *Phakopsora pachyrhizi,* in South America. *Plant Pathology*, 64 (3), 729-737.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596 (7873), 583-589.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30 (4), 772-780.

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome research*, 27 (5), 722-736.

Kües, U., James, T. Y., & Heitman, J. (2011). Mating type in basidiomycetes: unipolar, bipolar, and tetrapolar patterns of sexuality. In *Evolution of fungi and fungal-like organisms*, 97-160. Springer, Berlin, Heidelberg.

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. *Molecular biology and evolution*, 35 (6), 1547.

Letunic, I., & Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic acids research*, 49 (W1), 293-296.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34 (18), 3094-3100.

Li, F., Upadhyaya, N. M., Sperschneider, J., Matny, O., Nguyen-Phuc, H., Mago, R., ... & Figueroa, M. (2019). Emergence of the Ug99 lineage of the wheat stem rust pathogen through somatic hybridization. *Nature Communications*, 10 (1), 1-15.

Loehrer, M., Vogel, A., Huettel, B., Reinhardt, R., Benes, V., Duplessis, S., ... & Schaffrath, U. (2014). On the current status of *Phakopsora pachyrhizi* genome sequencing. *Frontiers in Plant Science*, 5, 377.

McDonald, B. A., & Linde, C. (2002). Pathogen population genetics, evolutionary potential, and durable resistance. *Annual review of phytopathology*, 40 (1), 349-379.

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., & Steinegger, M. (2022). ColabFold: making protein folding accessible to all. *Nature Methods,* 1-4.

Morgulis, A., Coulouris, G., Raytselis, Y., Madden, T. L., Agarwala, R., Schaffer, A. A., ... & An, P. (2008). BLAST+: architecture and applications. *Bioinformatics*, 24 (764), 1757-1764.

Ni, M., Feretzaki, M., Sun, S., Wang, X., & Heitman, J. (2011). Sex in fungi. *Annual review of genetics*, 45, 405.

Ono, Y., Buriticá, P., & Hennen, J. F. (1992). Delimitation of *Phakopsora*, *Physopella* and *Cerotelium* and their species on Leguminosae. *Mycological Research*, 96 (10), 825-850.

Paoletti, M., Rydholm, C., Schwier, E. U., Anderson, M. J., Szakacs, G., Lutzoni, F., ... & Dyer, P. S. (2005). Evidence for sexuality in the opportunistic fungal pathogen *Aspergillus fumigatus*. *Current Biology*, 15 (13), 1242-1248.

Pendleton, A. L., Smith, K. E., Feau, N., Martin, F. M., Grigoriev, I. V., Hamelin, R., ... & Davis, J. M. (2014). Duplications and losses in gene families of rust pathogens highlight putative effectors. *Frontiers in plant science*, 5, 299.

Prabantu, V. M., Naveenkumar, N., & Srinivasan, N. (2021). Influence of disease-causing mutations on protein structural networks. *Frontiers in Molecular Biosciences,* 7, 620554.

Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Höhna, S., ... & Huelsenbeck, J. P. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic biology*, 61 (3), 539-542.

Rydholm, C., Szakacs, G., & Lutzoni, F. (2006). Low genetic variation and no detectable population structure in *Aspergillus fumigatus* compared to closely related *Neosartorya* species. *Eukaryotic cell*, 5 (4), 650-657.

Schöneberg, T., Schulz, A., Biebermann, H., Hermsdorf, T., Römpler, H., & Sangkuhl, K. (2004). Mutant G-protein-coupled receptors as a cause of human diseases. *Pharmacology & therapeutics*, 104 (3), 173-206.

Schwessinger, B., & Rathjen, J. P. (2017). Extraction of high molecular weight DNA from fungal rust spores for long read sequencing. *Wheat Rust Diseases: Methods and Protocols*, 49-57.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31 (19), 3210-3212.

Slaminko, T. L., Miles, M. R., Frederick, R. D., Bonde, M. R., & Hartman, G. L. (2008). New legume hosts of *Phakopsora pachyrhizi* based on greenhouse evaluations. *Plant Disease*, 92 (5), 767-771.

Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics*, 14 (2), 178-192.

Twizeyimana, M., Ojiambo, P. S., Haudenshield, J. S., Caetano‐Anollés, G., Pedley, K. F., Bandyopadhyay, R., & Hartman, G. L. (2011). Genetic structure and diversity of *Phakopsora pachyrhizi* isolates from soybean. *Plant Pathology*, 60 (4), 719-729.

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P. C., Hall, R. J., Concepcion, G. T., ... & Hunkapiller, M. W. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, 37 (10), 1155-1162.

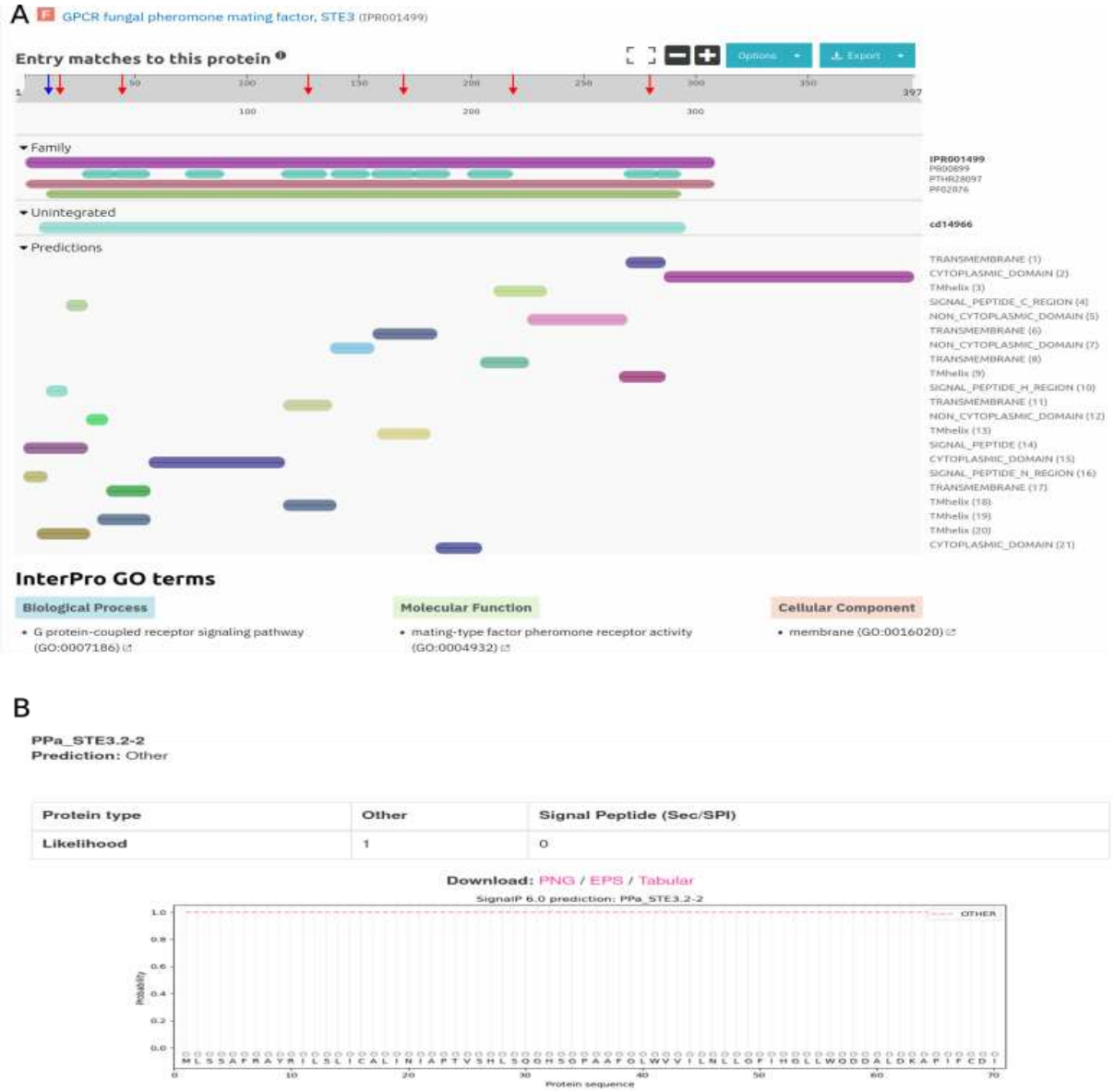
Wu, J. Q., Song, L., Ding, Y., Dong, C., Hasan, M., & Park, R. F. (2021). A chromosome-scale assembly of the wheat leaf rust pathogen *Puccinia triticina* provides insights into structural variations and genetic relationships with haplotype resolution. *Frontiers in Microbiology*, 2180.

Yamaoka, Y. (2014). Recent outbreaks of rust diseases and the importance of basic biological research for controlling rusts. *Journal of general plant pathology*, 80 (5), 375-388.

Yorinori, J. T., Paiva, W. M., Frederick, R. D., Costamilan, L. M., Bertagnolli, P. F., Hartman, G. E., ... & Nunes Jr, J. (2005). Epidemics of soybean rust (*Phakopsora pachyrhizi*) in Brazil and Paraguay from 2001 to 2003. *Plant Disease*, 89 (6), 675-677.

Zhang, J. (2003). Evolution by gene duplication: an update. *Trends in ecology & evolution*, 18 (6), 292-298.

## SUPPLEMENTARY MATERIAL




**Supplementary Figure 1:** Predicted domains of the pheromone receptor protein STE3.2-2 of *P. pachyrhizi* using Interproscan and SignalP6.0. A) Interproscan result indicates six transmembrane domains and a signal peptide domain (red arrows indicate regions predicted as transmembrane domains and blue arrow the predicted signal peptide region). B) SignalP6.0 did not predict the signal peptide region at the receptor protein.

**Supplementary Table 1**: Database used in phylogenetic analysis and *P. pachyrhizi* mating-type genes annotated from Mycocosm (JGI) and NCBI using PPUFV02 genome as a reference.

| Specie/genome | Protein | Access (JGI) |
|---|---|---|
| *Cronartium quercuum* f. sp. *fusiforme* G11 v1.0 | STE3.2-1 | Croqu1_66280 |
| *C. quercuum* f. sp. *fusiforme* G11 v1.0 | STE3.2-2 | Croqu1_666007 |
| *C. quercuum* f. sp. *fusiforme* G11 v1.0 | STE3.2-4 | Croqu1_54044 |
| *Melampsora larici-populina* v2.0 | STE3.2-2 | Mellp2_3_123740 |
| *M. larici-populina* v2.0 | STE3.2-1 | Mellp2_3_123975 |
| *M. larici-populina* v2.0 | STE3.2-4 | Mellp2_3_86096 |
| *M. larici-populina* v2.0 | STE3.2-3 | Mellp2_3_73569 |
| *Puccinia graminis* f. sp. *tritici* v2.0 | STE3.2-1 | Pucgr2_275 |
| *P. graminis* f. sp. *tritici* v2.0 | STE3.2-2 | Pucgr2_3045 |
| *P. graminis* f. sp. *tritici* v2.0 | STE3.2-3 | Pucgr2_4667 |
| *P. striiformis* f. sp. *tritici* 104E137A- | STE3.2-1 | Pucst_PST78_1_1854 |
| *P. striiformis* f. sp. *tritici* 104E137A- | STE3.2-3 | Pucst_PST78_1_6202 |
| *P. striiformis* f. sp. *tritici* 104E137A- | STE3.2-2 | Pucst_PST78_1_6788 |
| *P. triticina* 1-1 BBBD Race 1 | STE3.2-2 | Puctr1_11043 |
| *P triticina* 1-1 BBBD Race 1 | STE3.2-1 | Puctr1_3271 |
| *P. triticina* 1-1 BBBD Race 1 | STE3.2-3 | Puctr1_4853 |
| *C. quercuum* f. sp. *fusiforme* G11 v1.0 | HD2-1 | Croqu1_661468 |
| *C. quercuum* f. sp. *fusiforme* G11 v1.0 | HD1-1 | Croqu1_661465 |
| *M. larici-populina* v2.0 | HD2-1 | Mellp2_3_124184 |
| *M. larici-populina* v2.0 | HD1-1 | Mellp2_3_90168 |
| *P. graminis* f. sp. *tritici* v2.0 | HD1-1 | Pucgr2_2397 |
| *P. graminis* f. sp. *tritici* v2.0 | HD2-1 | Pucgr2_2396 |
| *P. striiformis* f. sp. *tritici* 104E137A | HD1-1 | Pucstr1_17785 |
| *P. striiformis* f. sp. *tritici* 104E137A | HD2-1 | Pucstr1_17786 |
| *P. triticina* 1-1 BBBD Race 1 | HD1-1 | Puctr1_12911 |

| | | |
|---|---|---|
| *P. triticina* 1-1 BBBD Race 1 | HD2-1 | Puctr1_12912 |
| *Phakopsora pachyrhizi* UFV02 v2.1 | HD1-1 | PPUFV02_4444554 |
| *P. pachyrhizi* UFV02 v2.1 | HD2-1 | PPUFV02_4444553 |
| *P. pachyrhizi* UFV02 v2.1 | STE3.2-1-1 | 4206263 |
| *P. pachyrhizi* UFV02 v2.1 | STE3.2-1-2 | 1579316 |
| *P. pachyrhizi* UFV02 v2.1 | STE3.2-2 | 3036970 |
| *P. pachyrhizi* UFV02 v2.1 | STE3.2-3 | 3374518 |
| *P. pachyrhizi* UFV02 v2.1 | HD1-1 | 4444554 |
| *P. pachyrhizi* UFV02 v2.1 | HD1-2 | 717051 |
| *P. pachyrhizi* UFV02 v2.1 | HD2-1 | 4444553 |
| *P. pachyrhizi* UFV02 v2.1 | HD2-2 | 4444555 |
| *P. pachyrhizi* UFV02 v2.1 | Mfa2 | 4444556 |
| *P. pachyrhizi* UFV02 v2.1 | Mfa3-Copy-1 | 1585066 |
| *P. pachyrhizi* UFV02 v2.1 | Mfa3-Copy-2 | 1143471 |

| **Specie/genome** | **Protein** | **Access (NCBI)** |
|---|---|---|
| *Austropuccinia psidii* MF-1 | STE3.2-1 | MF1_contig21007 |
| *A. psidii* MF-1 | STE3.2-2 | MBW0480530.1 |
| *A. psidii* MF-1 | STE3.2-3 | MBW0490563.1 |
| *Sphaerophragmium acaciae* | STE3.2-1 | USF89503.1 |
| *S. acaciae* | STE3.2-2 | USF89502.1 |
| *S. acaciae* | STE3.2-3 | USF89504.1 |
| *Ustilago hordei* | UhPra1 | Q99063.1 |
| *A. psidii* MF-1 | HD1-1 | MBW0530680.1 |
| *A. psidii* MF-1 | HD2-1 | MBW0519276.1 |
| *S. acaciae* | HD1-1 | USF89489.1 |
| *S. acaciae* | HD2-1 | 1032564_edges=1144221..3939684 |

**Supplementary data:**

Mating-type peptides and proteins predicted at *Phakopsora* sp., *P. pachyrhizi,* and *P. meibomiae* genomes.

>P.Pachy-PPUFV02_Ste3.1-1-1_ctg_164
MDNVNLLATVAYSTGCLIVAVLAAVPSAWFMYHGQSAPASLGIWVSISNLVHGINSTVWRVDV
VNRAPVWCDIASKIILIYSTGSACSCLCIAKFLAYALSPNAVNLSYDVRKWVNIRNYIFSFGFPI
GMIPFHYLYSPNRFILVRTIGCEASYVVTWTSVFFFIIWAPIFGTITAVYTVYVAYKLYEQKSKGI
FKKLKNSKVPTVRLAWLCIVYTAIALPMLIYYAIVLLQNGGYSALILSEIRSQASKIEYSSIQSSP
GVVDSLSIIGGAIYVAFFTFSQDLRKAYCKTFATIVKYISKLWKRSKVYKKRSTIENVPNMQTV
ALTISDTQFDDSLELMERQQVKSKALKFSNDKNRAFSLRSFSKKIHNHLGLSLITSELTASSPSS
RLQTSLEPLKPNNQSP*
>P.Pachy-PPUFV02_STE3.2-2_ctg_6
MLSSAFRAYRILSLICALINIAPTVSHLSQGHSGPAAFGLWVVILNLLGFIHGLLWQDDALDKA
PIFCDISAKIGLIGPLGLLMANCCIIRYLAQIVTPSNSIETSYEARKRMYKDYAMAFGFPVVIAIA
SVVFQVARYEVEILVGCSNVSVLGWPTLIILIVWSPIICGISCGYALYVLYWLVQQHKNLRQLVA
KSSTPLNMSRFVRMCALAATYLCISAPYTIAGTVTTLYDIGPFIPWKSWSYIHNDDNQLSNVRN
NPVYRLNVRDWLSITAGLTVFFFFSFAKESLDVYRKVGKALSLKSFLQRFSGSLTVIWPFNRLI
NNFSENKEPENKYCVQKELFTFNPSKTEINYHDQVILPLSASRSRVEKKMQLVDQFSLTRPIFID
KSALNKQNP*
>P.Pachy-PPUFV02_STE3.2-3_ctg_257
MTSDNIIRAYLALSFLCGIINIPPTLLHVVQGHSGPGAFGVWAVVLNLLAFVNGLLWRHDAVD
RAPIFCDLSSHVSQVGPLGLLIGNCCIMRYLAKIMTPNTTVEDQSEKRRRAVFDYSISFGFPAVL
AALSTIFQVARYQVNRYAGCSSASALVWPTFVLSVIWPPIFCGIACGYSFYVSYWLLQRHREIK
KLVEFSHTPLNVSRFIRMGALSTTYFFISTPCAVYGTLETLAATGPFVPWFSWATVHNEDNNLSI
IRQYPLYQYQLRDWLPIAAGLIVLMFFSFGAESQAFYKKLGLAALRIFTTRQKHMNKSFCISVD
VARSHCFKTEPKAGHNLIKSTIVAHKSNTTLSEQDPGSWKENTSGIC*
>P.Pachy-PPUFV02_HD-1-1_ctg262
MVPISLEFLRLQDRCFATSPDNISALGSLATENQRLMEQAEIASRSDMMTTSEAIACASIASNIV
QLSSCVIESNYAIQDLCTEFLKKTRFFSPNLNTYTSSIEPSSTIDRCPKPINLTQHGLRKWCKSHL
SYLFPTNKQVAEFSMISEMSESQVSSWFRNARTRSGWSKLFSDKVHVNRDPSRLDIVFQDYNS
YILSHCHSDICERLSYDVHFRLVDKVWRWFHTDKNCKKLEGSGSVRPWVKTVLFKALDKAK
ESCISDFNSKDAHPFSGFNEPPTQSFFTKKYPHSDSEDTASESSSSTNSRCSTLSCLSCELVSGSS
SGQPSSSSSTLHHVTTPSSTSSDYLPASNSSNFQSLSPSRVSSDLSTSDSHDYQSPPDLTFPLQSSP
PSLLTSHTLNPVEIINGLDALFKLDCSSGSPHQLPPPHSKYPEASTLPSDFFLGSLREFALPSSSST
NSQETAELESGQFDDASEN

>P.Pachy-PPUFV02_HD-2-1_ctg262
METTHFIRSDGKDALIYWENIRLHAVSILDYCDKFSVPEKFSHFNNKPIPSPMPPLTFPSLQFMR
TRLECLNLSRTLHVQVAGMLKQAISAIDKALHESYQNLLPLMQAWAPPHASHSWSTNDYAVA
VQNVLRGLRSDSVEKLWTSLLLNLPTILGATTGETKRPNACPWSMTPLHPRSTYKNGRPLDD
KQLIATALNLTREQVNRWFCNARARKKPYQKMKAKKTVSSILSSAKSSPTFETSNQAQP
>P.Pachy-PPUFV02_HD-1-2_ctg269
MESICLEFSGLHDRYFATAPDDTSTLGFLSEKLQHARERVKAASRSGLLTPLEATTCASIASNIE
KLSSCVIKSSHAIEDLCAEFSKKSHIFSPSLGTPSEPSLAVASCTSRTSSNQGKLKEWCRSHFSYL
FPTDSEVAKLSTISKMSENQVSSWFRNARTRSGWSKLYADKIHVNRDPIRLEIVFQDYHSYILL
HCQSDIREHISRDEHFRLVDKVWHWFQIDKKCKKPEDSGAVRPWVKTVLFEALQSSKESCNS
DIKLKEAPSFSGFGELPQISDQPFFPKLNPSSDIEETASESYSSTNYRCSTISYSSCESVSGSSSDQ
PMSSTSHHVTTPSSSILDFTPSFNSYSSQPSSPSIVSSDLSTPESQDYPSPPDLNCLHQLSPPSLFSS
QTSNPVEIINGLDALFQLDYSSDFFLDSLRDFAMPSSSSTNSQETSEFESGQFDDASED
>P.Pachy-PPUFV02_HD-2-2_ctg269
MHTTVMMKTSAVNEVTSAWDNIRQPAALILDACKTYPMSKHPVHDRIIYEAATPLPPLILPSLP
CFKSQLESLCLPSTVQAKVEGILMETLSSIDVALQESYCRHLPLMVAWADPQASQSRSMEDYA
QALQNMLWELRSDSVGKLWSSLMLHLRTLQGAARADGTSYEETQPSSSGSSHPGKISGQGRP
CAFTKEQTMVLRDLLAHDDRYSADDKQLIATALNLTREQVNRWFCNARARKKPYQKLRGYK
PASSILSNTTSYANSEPLLKVPQLPSTFQTYHFQASLSQHTIDPSPTPNYSPPDKSFLGSRHHLRN
LSVGRVDFSFANGPLHNYSNQSNQSAVTVQ*
>P.Pachy-PPUFV02_mfa3
MSFTKDTSSSIPLKDENGQQWGHGSHLCVIEQEWGNGSHMCITSQEWGNGSHLCVASNKGDS
GHFLNQEWGNGSHLCIVN
>P.Pachy-PPUFV02_mfa2_ctg06
MNHSLSNPSTNKELGGSNHNCIVSKELGGSNHNCVISREIGGSSHNCIIAKELGGSNHNCIIAEK
DNFSSHESTKELGGSNHYCIVS
>P.meib_STE3.1-1-1_MG19.3.5-ptg008
MADVNLLTTTAYSISCLLISILSAIPATWFLYHGQSAPASLGIWVTISNLIHGIDSTIWKIDVINRA
PIWCDFASKIILIYSTGSACSCLCIAKFLAYALSPDVVNLSYDVRKRVNMRNYIFSFGFPIGMIPF
HYLYSPNRFILVRTIGCEASYTITWTSLFFLIIWAPIIGSITAIYSAYVAYRLYEQKKQGTFKNLKN
SKVPTVRLAWLCIIYTTIALPMMFYYAFVLLTNGDYYPLILSEIRSKASEIEYSDLQSSPGFSDSL
SIFGGLVYVGFFTFSQDLRKAYSKNFLRILKVLTKLWNSKEIFYKKSWKMDSIINLNDINLNESE
TRFEDSLELMETQQTKSKL*
>P.meib_STE3.2-2_MG19.3.5-ptg091
MLSPTFEAYRILSLMCALINIAPTISHLSQGHSGPAAFGLWVVILNLLGFVHGIMWKDDALDRA
PIFCDISAKVKVGPLGLLMANCCIIRYLAQIVTPSSSVETLYEARIRMYKDYAMAFGFPVVVAII
SVVFQVARYEVEILVGCSNVSVLGWPTLIIFIVWSPIVCGISCGYAIYVLYWLVQQHKNLRQLVA

KSSTPLSMSRFVRMCALAATYLCVSAPYTIAGTVTTLYDIGPFIPWNSWSYIHNDDNQLSNVR
NNPIYHLNVRDWISITAGLTVFFFFSFAKESLEVYTKLGKLLRFNHLYKHLLASAKNKWPRNW
WSNGFSGKKKYTK*

>P.meib_STE3.2-3_MG19.3.5-ptg053

MTSDNVTNSYLTLSFLCAVINIPPTLLHLAQGHSGPGAFGSWAVALNLLAFVNGVLWRHDAVD
RAPIFCDFSSHLSQVGPLGLLISNCCIMRYLARIMTPNLAIEDRSEKRRRVFLDYGTSFGFPAVL
AALSTVFQVARYQVNRFAGCSSASALVWPTFLLSVIWPPIFCGIACGYSSVYVLYWLSRRHREI
KKLVECSQTPLNVSRFARMTALSATYFFVSTPCAVYGTLETLAATGPYVPWFSWSTVHNEDNN
LSIVRQYPLYQYQLRDWLPIAAGLIVLVFFSLGAESQAFYKKIGFAALKILPSIRNDKSKDLCIN
VDIAQSYSLKTCVK*

>P.meib_Hd1-1-MG19.3.5-ptg017

MASLAQHLSKLQDRCFTASPNDSINVARLTEELREVAFNLEEVSDIDSLPESELEASVSMANNIR
ELSLCVIESSNAFEDLVKSFSKKLSLFSLEQAPQALETVESEPTRKRGEFSKSSGPSQVILKEWC
KGHLQYLFPTHPQVKELAVTSNMSESQVNSWFRNARTRSGWSKLFANKSIVNRDPNRMKIVF
QDYHSYMLSNHKLDARPHLSSDEGFRLAHKVWCWLQTDKKNQKPEGSSKVKPWVKTALFH
AIGKARAGCKNSSKPKEGHSIPSGSAQSIFSVYLNSECEDSISESSSSYNSRCATFSFPSIASSFTS
SFQPINSKTTSSQATLTSSTFSGSLHPSKSSGVQSSTPSTLVPCICTPESNGLLSPPSSNISYQSARS
TCPSEIINSPEFFAGLNDLFLRKNSPPGDFFTAPLCQRIVSHDLPSHVSNKLPPTPGPISSRQSSPK
TVPPNITPPAKLFGDLSMLFRLGGLPSNPFQNPSAFDGFFTAPYCPSTVSSDLSSLDSYDLPSPSG
SSGHYNTHPSAMINKTKTPIENLNDLGALFQPDSLSSSPFQNSSQGSDFLTAPSLPNGSSFSSLFS
FILPSSNSFFDQEDSGIEPGQFDDASED*

>P.meib_Hd2-1-MG19.3.5-ptg017

MSFSCEDANPWNYSLRSAIIIQDTLRSFRNHETSKICENKKSSSIPMLSFPKAFCFGRVLEDLGL
PVAVRSKVERVLEETISAIDNTLQECYQECVPQMRVLNVPCGGRSWSQEDYAMAVQNTLLRL
RSGFVYRLWNCLLLSIRQGAAWLGNRKPGASLQSRLREDDDSSSRAKVGRPCLFSKEQTMVL
RALLAHDDRYSPDEKQLIATALNLTREQVSRWFCNARARKKPYQRTKSKKSVSSVLRRIESSPA
FEPQTQASRQCSISLSPQYQSSVLPSTTESSHFYFGGSPDYSAQEKSTMVFDDRLNFLFGDRMD
FSNLKRDTNLNSYFQNLINPLAVAVQ*

>P.meib_Hd1-2-MG19.3.5-ptg039

MTSLAQHLSKLQYRCFTASPNDSINVAHLTEGLREVAFKLEEVSDIGSLPKSELEASVSIANNIR
ELSSCVIESSNALEELIESFSQKLSVFSLEQAPQAVETLEPEPNRKRGECPKSSSPSQIILKQWCTS
HLRYLFPTHTQVTELAVTSNMSESQVNSWFRNARTRSGWSKLFADKSRVNRDPDRMEIVFQD
HRSYMLSSDKSDVQACLSSDESLRLADKVWRWFQTDKKNPKPEESDNVRPWVKTVLFNAIE
KVQASSKNSSPKETRLISGLEAIPSQSAQSTLSTYPPSESEESTLESSSNKSRCSSFSFPLFESHLTS
SSFQQTIPKTPSSRATSPTSTCDHLHLSDSSLVQSTSPSMLSREASTPDLHDLPSPPGSIRNCQSIQ
SIVPPKSIAPSKIFSALDYDLVQLESFQSNPFQNSWPSSDFFTASSFSSVISSDIYSPDLHNLLSPAG
SSNSPQSISSASPGKTTTLSENFSGLNALLELDDFSSNIFKNPWPPSNFFTVPSCQSIVSSGIYSPD

SHDLPPPPGSSSFIQSDPSEVLPKIKTSTENPNDSGAIFQLDDLPSEPFPNSPPPSDFFVTPTLSSVS
SFNSLFNFTIPHQDLIFDQTNLSIESGQFDDAPDE*
>P.meib_Hd2-2-MG19.3.5-ptg039
MGFSFEEANSWNYSLRSAVIIQDTLRSYRNNVTLKICENKNSSSIPLLSFPKAFCFGRVFEDLGL
PVALRSKVERVLEETISAIDITLQECYQECVPQMRVLNVPYGVRSWSREDYAMAVQNTLLSLR
SGYIYQLWNCLLLSIRQGSTWLENRQPGASLQSLLRGSSSSSRMKVGRPCLFSKEQTMVLRAL
LAHEDRYSPEDKQFIASSLNLTRKQVNRWFCNARARKKPYQRNKGRKSVSSVLGRIESSQDSE
PDPQVSKPCSISSSPPYQDSLLPTTAESTPASFSLPDYTTQEKSQKVFEPLNFLLGDRMDFISLNC
DTNLNSYFQNLVNPSSVVVQ*
>P.meib_Mfa2-1-19.3.5-ctg91-4550kb
MIQSFTKPSINQEIGGSNHYCVISQELGGSNHYCVLSKEIGGSNHNCIIAQELGGSNHYDVTSK
EIGGTNHNCIILHELGGSNC*
>P.meib_Mfa2-2-19.3.5-ctg91-4551kb
MIQSFTKPSINQEIGGSNHYCVISQELGGSNHYCVLSKEIGGSNHNCIIAQELGGSNHYGVTSK
EIGGTNHNCIILHELGGSNC*
>P.meib_Mfa2-3-19.3.5-ctg91-4678kb
MIQSFTKPSINQEIGGSNHYCVISQELGGSNHYCVLSKEIGGSNHNCIIAQELGGSNHYGVTSK
EIGGTNHNCIILHELGGSNC*
>P.meib_Mfa2-4-19.3.5-ctg91-4679kb
MIQSFTKPSINQEIGGSNHYCVISQELGGSNHYCVLSKEIGGSNHNCIIAQELGGSNHYGVTSK
EIGGTNHNCIILHELGGSNC*
>P.meib_Mfa2-5-19.3.5-ptg000342-76kb
MMQSFTKPSINQEIGGSNHYCVISQELGGSNHYCVLSKEIGGTNHNCIIAQELGGSIHYGVTSR
EIGGTNHNCIILQELGGSNC*
>P.meib_Mfa2-6-19.3.5-ptg000342-133kb-trunc
MIQQPSVKQEIGGSNHYCVISQ*LGGSNHYCVLSKEIGGTNHNCIIAQELRGSNHYGVTSKEIG
GTNHNCIVLQELGGSNCLCT*
>P.meib_Mfa3-1-19.3.5-ctg53-4583kb-ORF1
MNLITLNDYSPLSASAHVNSQEWGNGSHLCTIEQEWGNGSRSCVVSQEWGNGSRSCVVSQE
WGNGSHMCVSTNNAGDEPIDT*
>P.meib_Mfa3-2-19.3.5-ctg53-4631kb-ORF1
MNLITLNDYSPLSASAHVNSQEWGNGSHLCTIEQEWGNGSRSCVVSQEWGNGSRSCVVSQE
WGNGSHMCVSTNNAGDEPIDT*
>P.meib_Mfa3-3-19.3.5-ctg53-4659kb-ORF1
MNLITLNDYSPLSASAHVNSQEWGNGSHLCTIEQEWGNGSRSCVVSQEWGNGSRSCVVSQE
WGNGSHMCVSTNNAGDEPIDT*
>P.meib_Mfa3-4-19.3.5-ctg53-4704kb-ORF1

MNLITLNDYSPLSASAHVNSQEWGNGSHLCTIEQEWGNGSRSCVVSQEWGNGSRSCVVSQE
WGNGSHMCVSTNNAGDEPIDT*
>P.sp_STE3.2-1-1-MG19.8.2-ptg032
MGDVNLLATTAYSTSCLLISFLASIPASWFLYHGQSAPASLGIWVTVSNLIDSTVWKVDVINRA
PVWCDIASKIILIYSTGSACSCLCISKFLAYALSPDAVNLSYDVRKRVNIRNYLFSFGFPIGMLPF
HYLYSPNRFVLVRTIGCEASYIVTWTCLFFFIIWAPIIGTITAVYAAYVAYRLYEQKKQGLFKKLK
NSKVPTVRLAWLCIAYTTIALPMMIYYVFVLLSNGGYSPLILSEIRSRANLKISPDFSDSFSIFGG
LVYVGFFTFSQDLRKAYLKIFLRISKFLSKLWKIKKIDYKESLNLDGIINLNAIALNESNTQFEDS
LELMESQQTKSKLQVYYYFL*
>P.sp_STE3.2-2-MG19.8.2-pgt019
MLSPAFEAYRILSLMCALINIAPTVSHLSQGHSGPAAFGLWVVILNLLGFVHGIIWQDDALDRA
PVFCDISAKVLVGPLGLLMANFCIIRYLAQIVTPSNSVETLYEARIRMYKDYAMAFGFPVVVAI
VSVVFQVARYEVEILVGCSNVSVLGWPTLIIFIVWSPILCGISCGYAIYVLYWLVQQHKNLRQLV
AKSSTPLNMSRFARMCALAATYLCVSAPYTIAGTVTTLYDIGPFIPWKSWSYIHNDDNQLSNV
RNNPIYHLNVRDWLSITAGLTIFFFFSFAKESLEVYTKLGKALRLNRLFKYLLAFLRNNWTRN
WLSYDVSEKQKYAK*
>P.sp_STE3.2-3-MG19.8.2-ptg036
MTSENITNTYLTLSFLCAIINIPPTLLHLAQGHSGPGAFGIWAVALNVLAFVNGVLWRHDAVDR
APIFCDFSSHLSQVGPLGLLIANCCIMRYLARIMTPNLAIEDQSEKRRRVFLDYATSFGFPAVLA
ALSTIFQVARYQVNRFAGCSSASALVWPTFVLSIIWPPIFCGIACGYSFYVLYWLLRRHREIKHL
VECSQTPLNVSRFARMAALSTTYFFVSTPCAVYGTLETLAATGPYVPWFSWSTVHNEDNNLSI
VRQYPLYQYQLRDWLPIAAGLIVLVFFSLGAESQAFYKKIGFAALRMLPPVRNNKRKDFCTD
MDVAQSFGFKKCVN*
>P.sp_Hd1-1-MG19.8.2-ptg033
MTSLAQHLSKLQDQCVTTSPNDFVNVSCLVEGLSEVAYKMEGATRLYPDPELEACDSMANNL
SELFLCVIESSSAFKELMKSFSEKLSLLSLEQPPLASETVESEPPRKFGEHSKSSSPSQLILKEWC
RSHMQYLFPTHHQISELALISNMSESQINSWFRNARTRSGWSKLFSNKSRVNRDPDRMEIVFQ
DYHSYMLSNQKLDVQAHFSNDEGFRLADKVWRWFQTDKKSEKPEESCKVRPWVRTVLFDAI
EEAQASSKISSEPKETRLISNLYTTPSKPDQSILPICTPSESGESTPESSSNKSQCSTFSFPLFKSDP
TSSSLQPTSPKTTSSHATSPSSACDYLHSSDSSSVQSSSPSILTSETSTPDLHDLPSPPGLSRCKPIQ
RTNPPDNFPDPDYDLLLHESFPINSFQISWPPSDFFSVLSCPRVLSSDMYSSDSHGLPSPPGSNSS
PHSISSASQSNVTNPIENFSGLNSFFGLDSLSSNLFKNPLPPSKPFTVPFCQSTISSEIYSHGLHELP
SPLGSSSSRQPNLSAVPSKITTPTENLNGLGDLFQLNHSPSNTFQNSSSPTDFFMVPSLPNGSPSY
SSSNFNLSCCNSIFNEANLDIESGQFDDAPEE*
>P.sp_Hd2-1-MG19.8.2-ptg033
MSFSRELADSWNHTLQSATIIHDTLRSFRNNAASKIGESKNPPSVPFLSFPKAFCFGRVFDDLGL
PFEARSKVSQVLEETIASIDFALHECYQECVPLMRTGDALCGNRQWSREDYAMAVQNTLLCL

RSGSIDRIWNCLVLSIRQGSHCLKNRHHGALIQSPCTSDSDKFTNASRPRLFSKDQTMVLRALL
DHDDRYSPEDKQLIATALNLTREQVNRWFCNARARKKPYQRTKCKKSVSAVLRRMESSPTFEP
HHQASPQSLMSSPPQYQCSISPSMTESSHSYSFGSPDCSPQGKSLMVLEPLDFLLGDRMDFSNL
NSDTNLNSYFQNLINPSAVAVQ*

>P.sp_Hd1-2-MG19.8.2-hap_069

MTSLTQHLSKLQEQYFTTSPDDHINIARLAKGLGEVACKVEEASNIDSLPESELGACAIMANNI
KELSLCVIESSDVFEELIKSFSEKLSLLPLGQAQQTLKTVESESTRKHEEFSKSSSSSQFILKECH
LQYLFPTHPQVVELAVISKMSESQVNSWFRNARTRSGWSKLFADKSSVNRDPARMEIVFQDY
QSYLLSNCNSDVRARLSSDEGFRLADKVLRWFQTDKKNQNPEESGKVRPWVRTVLFNAIEKA
QASSKNSSEPEEKRLISSLDEIPWRYAQSISSMYPPSESEESTPESSSNKSQCSTFSF

>P.sp_Hd2-2-MG19.8.2-hap_069

MSLSSGEANSWNHPLQSAIIIHDTLINYQHSAASKICENKNLTSMPLLSFPKAFCFERALEGMG
LPVALRSKFEGVLEETISAIDVSLQKCYQNFFRMLHVRDITCGDRSWTPEDYAMAVQNILLCLR
SRSLYQLWNCLVLSIRKGSHWLKNGHSGAPLQSNSWSSDPDKFSEAARPRLFNKDQTMVLRA
LLSHDDRYSPEDKQLIATALNLTREQVNRWFCNARARKKPYQRTKCKKTVSSVLRRMESSPAS
EPHPRHSPQCSISSSPQYQCSISPSTTGSSHSYSVRSPDFLPQEKSLKVFEPFDFLLCDRIDFSNLN
SDTNLNSYFQNLINPSAVAVQ*

>P.sp_Mfa2-1-19.8.2-ctg19-5323kb

MFHSLTNPFAKKELGGSNHNCVISKELGGINHYCAVSKEIGGTNHNCIISQELGGSNHYCIIPKE
IGGSNHYCVIAQELGGSNHKCTVV*

>P.sp_Mfa2-2-19.8.2-ctg19-5324kb

MFHSLTNPFAKKELGGSNHNCVISKELGGINHYCAVSKEIGGTNHNCIISQELGGSNHYCIIPKE
IGGSNHYCVIAQELGGSNHKCTVV*

>P.sp_Mfa2-3-19.8.2-ctg19-5827kb

MIHSLTNPSNNKELGGSNHNCVISKELGGSNHNCVVSKEIGGTNHSCMIAQQLGGSNHYCFIS
KELGGSNHYCVIAQEIGGSNH*

>P.sp_Mfa2-4-19.8.2-ctg19-5829kb

MIHSLTNPSNNKELGGSNHNCVISKELGGSNHNCVVSKEIGGTNHSCMIAQQLGGSNHYCFIS
KELGGSNHYCVIAQEIGGSNH*

>P.sp_Mfa3-1-19.8.2-ctg36-256kb

MTNFNASAQVNSQEWGNGSHLCTIEQEWGNGSHLCIIEKEWGNGSHICAISQQWGHGSHMC
VSTNNDGDELIDHQEWGNGSRYCLIY*

>P.sp_Mfa3-2-19.8.2-ctg36-kb-426kb

MTNFNASAQVNSQEWGNGSHLCTIEQEWGNGSHLCIIEKEWGNGSHICSISQQWGHGSHMC
VSTNNDGDELIDYQEWGNGSRYCLIY*

>P.sp_Mfa3-3-19.8.2-ctg36-kb-509kb
MTNFNASAQVNSQEWGNGSHLCTIEQEWGNGSHLCIIEKEWGNGSHICAISQQWGHGSHMC
VSTNNDGDELIDYQEWGNGSRYCLIY*

# CHAPTER 3

## A CHROMOSOME-LEVEL ASSEMBLY OF *Simplicillium lanosoniveum* GENOME SHED INSIGHTS ON MOLECULAR MECHANISMS ASSOCIATED WITH THE HYPERPARASITISM ON *Phakopsora pachyrhizi*

## ABSTRACT

*Simplicillium lanosoniveum* (Hypocreales) infects the aggressive Asian soybean rust pathogen *Phakopsora pachyrhizi*, colonizing uredinias and destroying its uredospores, being a potential agent for biological control. The antagonistic relationship between *S. lanosoniveum* and *P. pachyrhizi* is well described at a microscopic level, however the virulence factors involved in this process are still unknown. Here, it is report the first high-quality genome assembly and annotation of *S. lanosoniveum* isolate SL-UFV as well as the prediction of secondary metabolites clusters. Synteny analysis allowed the identification of patterns of mesosyntheny among species of the order Hypocreales and high synteny between *Simplicillium* spp. Using comparative genomics, 67 genes and seven carbohydrate-active enzymes (Cazymes) unique for *Simplicillium* spp. and seven other genes and Cazymes unique for *S. lanosoniveum* were identified. Eighteen chitinases and 69 proteases were identified in the *S. lanosoniveum* secretome. Thirty-two putative secondary metabolite clusters in the *S. lanosoniveum* genome were also predicted, including three that encode fungicide or fungistatic metabolites (Strobilurin, Acuelacin, and Squalestatin). The comparison of these three clusters with other Hypocreales fungi showed evidence of association with the *S. lanosoniveum* lifestyle including an additional Squalestatin gene cluster not present in other mycoparasite species analyzed. The results obtained in this work open new possibilities to understand the *Simplicillium lanosoniveum* mycoparasitism using the tools of molecular genetics.


Keywords: Mycoparasite. Genomics. Biological control. Squalestatin.

## 1. INTRODUCTION

Mycoparasitism and mycotrophy are antagonistic relationships between fungi that occur when a parasite or predator species uses another fungus as a host or prey to access their nutrients, establishing a biotrophic or necrotrophic interaction (Barnett, 1963; Karlsson et al., 2017). Phylogenetic studies support the hypothesis that the transition to the mycoparasitic lifestyle occurred several times over the Hypocreales order as an adaptative advantage to guarantee survival (Spatafora et al., 2007).

Hypocreales supports many fungi genera with mycoparasitic species such as *Trichoderma* spp*., Clonostachys* spp., *Escovopsis* spp*.,* and *Simplicillium* spp.. Among them, *Trichoderma* and *Clonostachys* are the most studied genera with 370 known and accepted species in the *Trichoderma* genus (https://trichoderma.info/trichoderma-taxonomy-2020/) (Cai & Druzhinina, 2021) and 48 identified species of *Clonostachys* spp. (indexfungorum.org; June, 2022), including fungi and oomycetes parasitic species, saprophytic species, and animal pathogenic species (Kubicek et al., 2019). On the other hand, *Escovopsis* and *Simplicillium* are considered smaller genera with 14 and 26 reported species, respectively (indexfungorum.org; June 2022), but with high scientific interest due to their possible use in biological control of agricultural pests, such as described for *E. weberi* in indirect control of leafcutter ants and *S. lanosoniveum* for Asian soybean rust control (Heine et al., 2018; Ward et al., 2012).

Comparative genomics among saprophytic and parasitic species of *Trichoderma* spp. with different parasitic behavior, associated with gene expression analysis, made it possible to identify genes possibly associated with parasitism mechanisms, such as antibiosis of *T. virens* and parasitism by contact employed by *T. atroviride* (Atanasova et al., 2013). These studies also supported the identification of differentially expressed genes and clusters during the antagonistic relationship that were associated with some species antagonism, such as secondary metabolites, GH16 beta-glucanases, proteases, small secreted proteins, cysteine-rich proteins highly expressed by *T. atroviride*, and gliotoxin precursors and glutathione necessary for gliotoxin biosynthesis by *T. virens* (Atanasova et al., 2013). Using a similar approach, some *Clonostachys* spp. genes expressed during mycoparasitism were also identified, such as the zearalenone hydrolase gene *zhd101* of *C. rosea* that was over expressed in the presence of the mycotoxin zearalenone produced by *Fusarium* spp. allowing the detoxification of the zearalenone toxin (Kosawang et al., 2014; Nygren et al., 2018). Although some antibiotic and fungicide compounds produced by *S. lanosoniveum* were previously

identified (Fukuda et al., 2014; Rukachaisirikul et al., 2019), none of these molecules was identified as one factor of virulence necessary for its mycoparasitism. On the contrary, the complete and annotated genome of *E. weberi* allowed the identification of genomic characteristics, genes, and gene clusters uniquely associated with its mycoparasitic relationship such as the absence of cellulases and the increased number of genes encoding chitinases, possibly associated with *E. weberi* mycoparasitism (de Man et al., 2016). Up until now, all genomic studies done with *Simplicillium* spp. are restricted to *S. aegoshimaense*, a mycoparasite of the wheat powdery mildew fungus, *Blumeria graminis* f. sp. *tritici* (Zhu et al., 2022).

*Simplicillium lanosoniveum* (SL) is a mycoparasite that infects different species of rust fungi (Pucciniales), such as *Phakopsora pachyrhizi* that causes the Asian soybean rust disease in soybean plants (Ward et al., 2011). The SL mycoparasitism process was very well described at the microscopic level by Gauthier et al. (2014). This mycoparasite wraps around urediniospores before colonizing them. An amorphous material that may have facilitated their attachment to urediniospores is observed between hyphae and urediniospores. Penetration occurs through germ pores within 24 h after inoculation, and hyphae of SL erupt from colonized urediniospores seven days after the inoculation. Degradation of germ pore material adjacent to hyphae of SL is often observed as well as disintegration of organelles within urediniospores. The cytoplasm of urediniospores of *P. pachyrhizi* aggregates, and cellular contents are unidentifiable 6 h after co-inoculation before significant amounts of hyphae of *S. lanosoniveum* are observed within urediniospores. These changes suggest that cell-wall degrading enzymes may aid the pathogen in entering germ pores. Other enzymes such as β-glucanase, chitinase, and protease may be associated with urediniospores degradation. However, the genomic features and virulence factors involved in the SL mycoparasitism are still unknown.

As demonstrated for *E. weberi* (de Man et al., 2016), genome characterization allows to understand better the ecological process and virulence factors possibly associated with the mycoparasitism. With this objective, it is reported here the assembly and annotation of a high-quality reference genome for *S. lanosoniveum* and the identification of genome features unique to this fungus that could be associated with its successful parasitism of rust fungi.

## 2. RESULTS

### 2.1 Mycoparasite isolation and identification

The first step in the analysis was to identify the mycoparasite associated with the uredinias in soybean plants infected with *P. pachyrhizi*. Superficial conidia of the mycoparasite in *P. pachyrhizi* infected uredinia were used for direct isolation of isolate SL-UFV that was maintained in Potato dextrose agar (PDA) medium. This isolate was inoculated in not infected *P. pachyrhizi* uredinias on detached soybean leaves maintained at Petri dishes and the mycoparasite signals (white mycelium) were observed 14 days later (Figure 1A), while no signal of the mycoparasite fungus was observed in the water-treated uredinias (Figure 1B). The aseptic culture in the PDA medium was used for DNA and RNA extraction used in whole genome sequencing and RNA-seq.

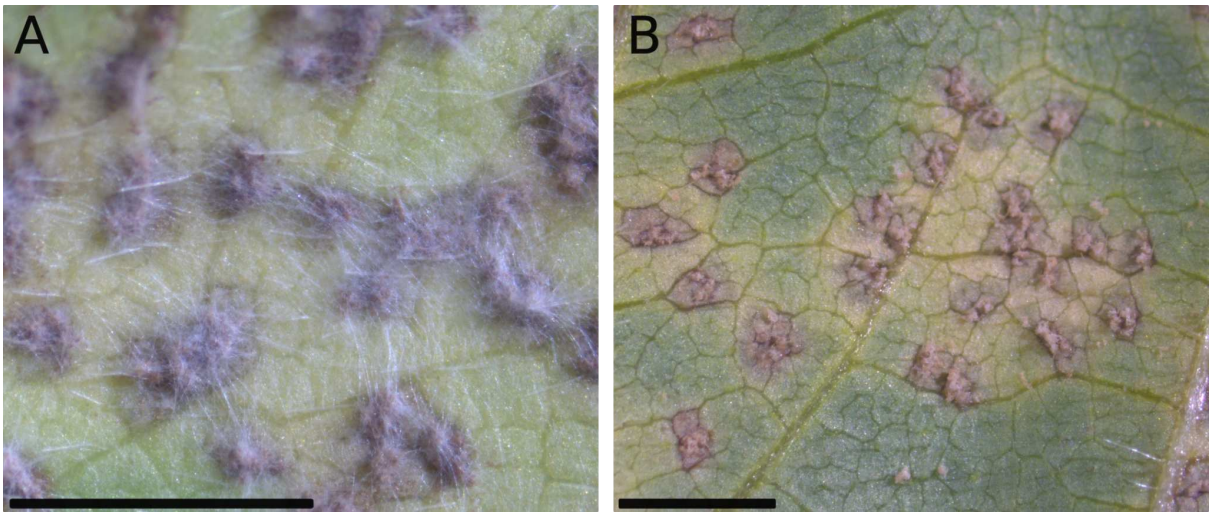



**Figure 1**: Soybean leaves inoculated with *P. pachyrhizi* and sprayed with *S. lanosoniveum* conidia suspension or water. Detached soybean leaves infected by *P. pachyrhizi* were inoculated with a SL-UFV conidia suspension (A) or sprayed with water (B) 14 days after the *P. pachyrhizi* inoculation and maintained in leaves in Petri dishes for 14 days. Photos were captured 14 days after the inoculation with *S. lanosoniveum* suspension (A) or water treatment (B). Black bars = 1 mM

Phylogenetic analyses were performed to complete the identification of the mycoparasite associated with *P. pachyrhizi* uredinias. For this, the genic regions *Internal transcribed spacer* (ITS), *Large subunit ribosomal ribonucleic acid* (LSU), and *Small subunit ribosomal ribonucleic acid* (SSU) of one *S. lanosoniveum* isolate (SL-UFV) were filtered from the genomic sequence obtained from total DNA extraction. The sequences were concatenated and aligned with the other 52 taxa of the *Simplicillium* genus, resulting in 2,296 characters with gaps (ITS: 565, SSU: 550, and LSU: 682), and used in phylogenetic

reconstruction analysis. The *P-value* of the Homogeneity Partition Test (*hompart* command in PAUP) was 0.16 and indicated the dataset SSU + ITS + LSU was congruent and suitable for the combined analysis. SL-UFV isolate clustered in a well-supported clade (posterior probability =. 1.0), in the phylogenetic tree, with two *S. lanosoniveum* isolates (Figure 2), confirming the previous morphological identification with the evidence of mycoparasitism on *P. pachyrhizi.*

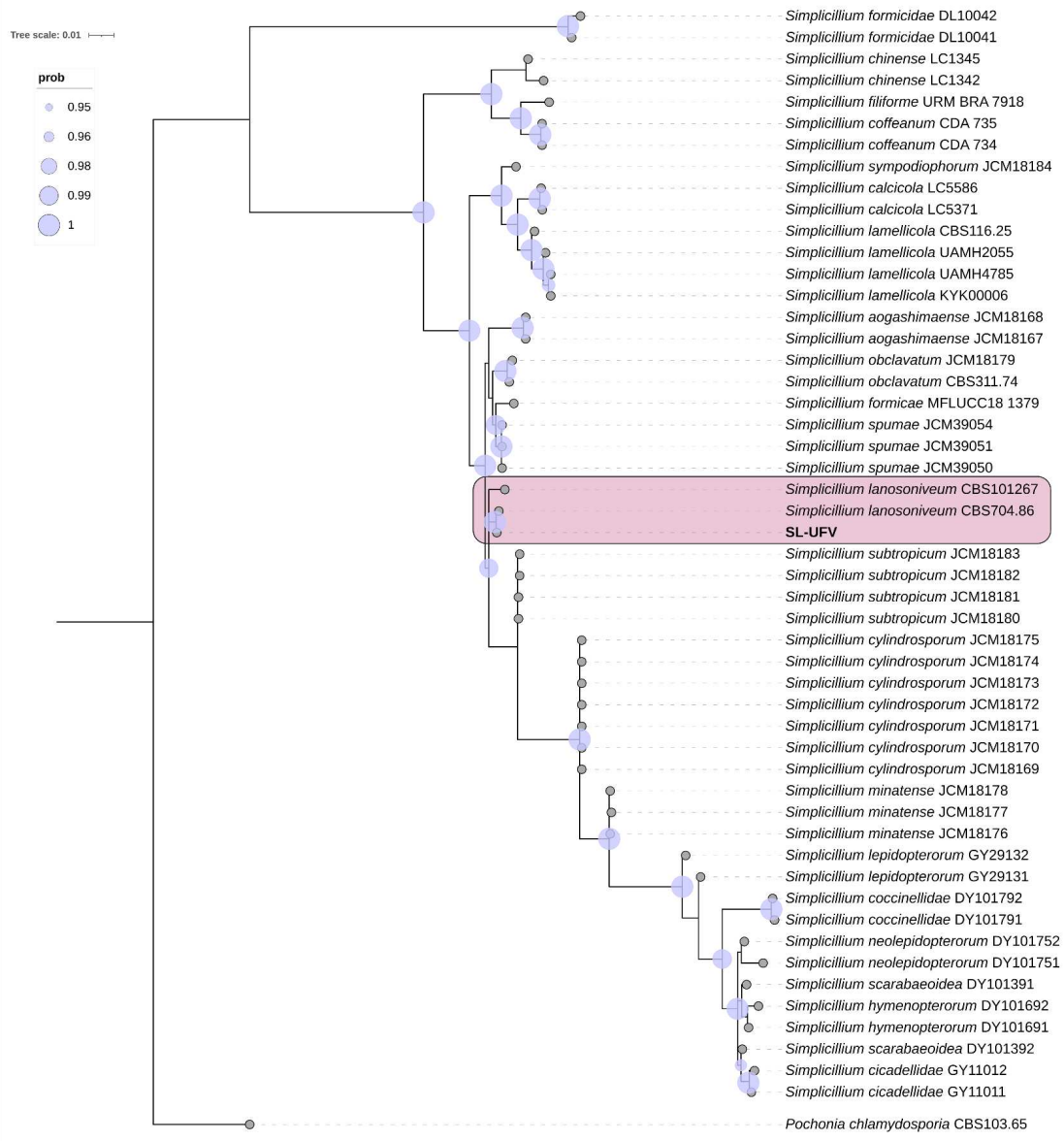



**Figure 2**: Consensus phylogram of the *Simplicillium* spp. for molecular identification of the isolate SL-UFV. The tree was constructed using Bayesian inference [implemented in MrBayes 3.1 (Huelsenbeck et al., 2001), with (SYM+I+G (ITS), GTR+G (LSU), and F81 (SSU) models, 20 million generations and samplefreq=5000] by the combination of three gene sequence alignments (ITS, LSU, and SSU) of 52 taxa of the *Simplicillium* genus. The tree likelihood log convergence was performed in TRACER v. 1.4.1 (Rambaut et al., 2018) and visualized and edited the tree using iTOL v.5. The tree was rooted with *Pochonia chlamydosporia* CBS 103.65 isolate.

### 2.2 Sequencing, genome assembly, and annotation

DNA sequencing was performed with the PacBio technology generating 918,088,293 bases distributed in 114,211.0 reads with an average HiFi read length of around 8.000 bp**.** The assembly of the genome of *S. lanosoniveum* was performed into eight contigs with a total length of 32.59 Mb and GC content of 49.86% (Table 1). Of the eight contigs, telomere sequences were identified in both extremities of four contigs, and in one extremity of three contigs (Figure S1). The annotation resulted in the identification of 6,779 genes, with an average gene length rounding 527 bp and encoding 6,695 proteins. The completeness of the genome assembly, generated with BUSCO v3.0.2 using the Hypocreales database (Simão et al., 2015), showed that most of the genes (97.9%) were identified as single-copy genes (Table 1), and the absence of duplicated genes.

**Table 1**: Descriptors of the *Simplicillium lanosoniveum* genome assembly, annotation, gene prediction, and completeness analysis performed with BUSCO.

| **Genome assembly** | | |
|---|---|---|
| Contigs number | 8 | |
| Max length (bp) | 5,989,139 | |
| Average contig (bp) | 3,621,325 | |
| N50 length (bp) | 4,711,838 | |
| Total length (bp) | 32,591,926 | |
| GC (%) | 49.86 | |
| **Completeness of the genome assembly (BUSCO)** | | |
| Complete (Single-copy) | 4,418 | 98.3% (97.9%) |
| Fragmented | 10 | 0.2% |
| Missing | 66 | 1.5% |
| Total | 4,494 | 100% |
| **Annotate Results** | | |
| Gene number | 6,779 | |
| Proteins Number | 6,695 | |
| Gene total length (bp) | 9,511,520 | |
| Gene average length (bp) | 527 | |

Functional annotation of the *S. lanosoniveum* genome resulted in 6,779 protein-encoding genes, which were grouped into gene ontology (GO) functional annotation. The GO-annotated genes showed the greatest number related to the Molecular Function category (48.5%), followed by biological processes (39%) and cellular components (12.5%). Of the 6,695 annotated proteins, 969 (14,5%) were predicted by SignalP to contain the signal peptide for the conventional secretion pathway, indicating they are secreted. Among proteins, it was identified 26 glucanases, 18 chitinases, and 69 proteases.

The gene relationships among *S. lanosoniveum* and other species: three mycoparasite (*T. atroviride, T. virens,* and *E. weberi*), one saprophytic (*T. reesei*), and two *Simplicillium* spp. (*S. aogashimaense* - from asymptomatic leaves of *Urochloa brizantha*; *Simplicillium* sp. - From seawater) (Figure 3A), as well as *S. lanosoniveum's* relationship with other *Simplicillium* species (Figure 3B) were analyzed by comparing the GO annotated genes present in each species. Of a total 1,195 GO annotated genes, 67 (5.6%) genes were specific of *S. lanosoniveum* and 7 genes (*ANP1 - Mannan polymerase II complex ANP1 subunit; CPC2 - ribosome-associated signaling scaffold, receptor of activated C kinase (RACK1) ortholog Cpc2; CSM1 - microtubule-site clamp monopolin complex subunit Csm1/Pcs1; GAR1 - H/ACA ribonucleoprotein complex subunit 1; SNF7 - Vacuolar-sorting protein SNF7; GOT2 - Aspartate aminotransferase; NUO-12 - Acyl carrier protein, mitochondrial*) were shared with other fungi mycoparasite species (*T. atroviride, T. virens* and *E. weberi*) (Figure 3A). The comparison among *Simplicillium* species demonstrated that *S. lanosoniveum* has more shared genes with the *Simplicillium* sp. than with *S. aogashimaense,* corresponding to 124 (10.4%) and 17 (1.42%) shared genes, respectively. The genic analysis using only *Simplicillium* species showed that 572 (59.9%) genes of a total of 955 were shared among the species (Figure 3B).

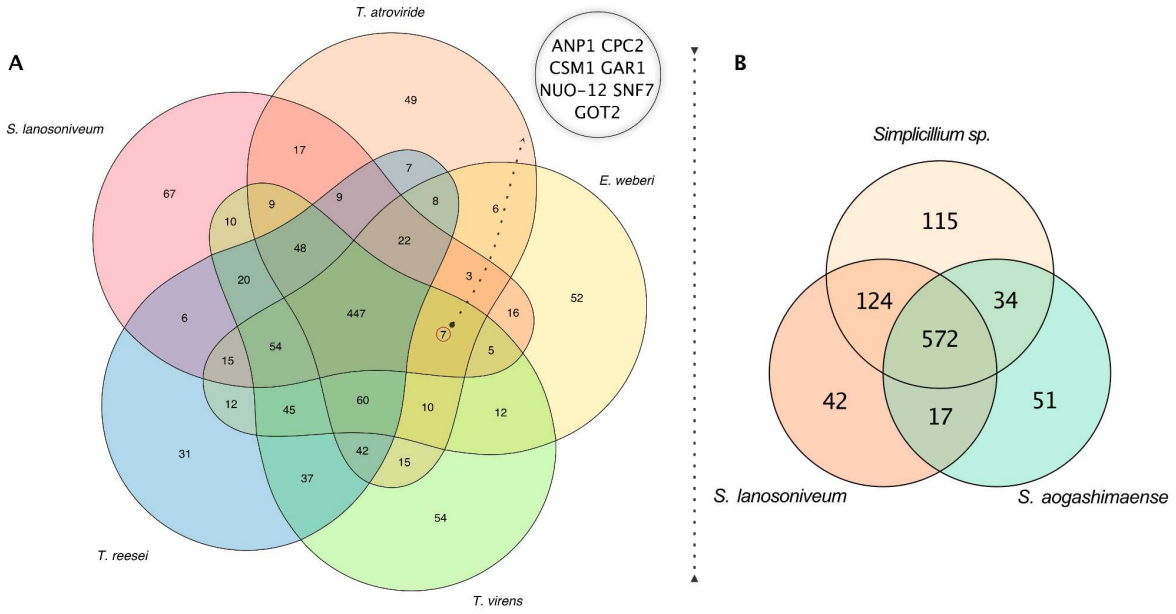



**Figure 3:** Venn Diagram demonstrating the gene content shared among *S. lanosoniveum, E. weberi,* and three *Trichoderma* species (*T. atroviride*, *T. virens,* and *T. reesei*) (A) and genes shared between *Simplicillium* spp. (B). Genes were assigned using genome annotations obtained from Funannotate. Each overlapping region corresponds to genes shared between species. The seven genes shared between mycoparasite species are highlighted with a black and red circle in A.

## 2.3 CDS sequence-based synteny among Hypocreales fungi

Syntenic blocks shared across fungi of the order Hypocreales (Figure 4A), including species of *Simplicillium* (*S. lanosoniveum, S. aogashimaense,* and *Simplicillium* sp.) (Figure 4B), were observed. These synteny results did not lead to identifying any pattern between the syntenic blocks of *S. lanosoniveum* and their position on the chromosomes of other species analyzed (*S. aogashimaense*, *Simplicillium* sp., *E. weberi*, *T. atroviride*, *T. virens,* and *T. reesei*). However, a clear mesosynteny among those species (Figure 4) was identified; in other words, the genes seem to be conserved within homologous chromosomes, but the order and orientation were randomized along the chromosome. The mesosyntenic blocks shared by *Simplicillium* sp. and *S. lanosoniveum* (Figure 4B) showed more similar patterns of order and orientation, demonstrating synteny occurrence between related species.

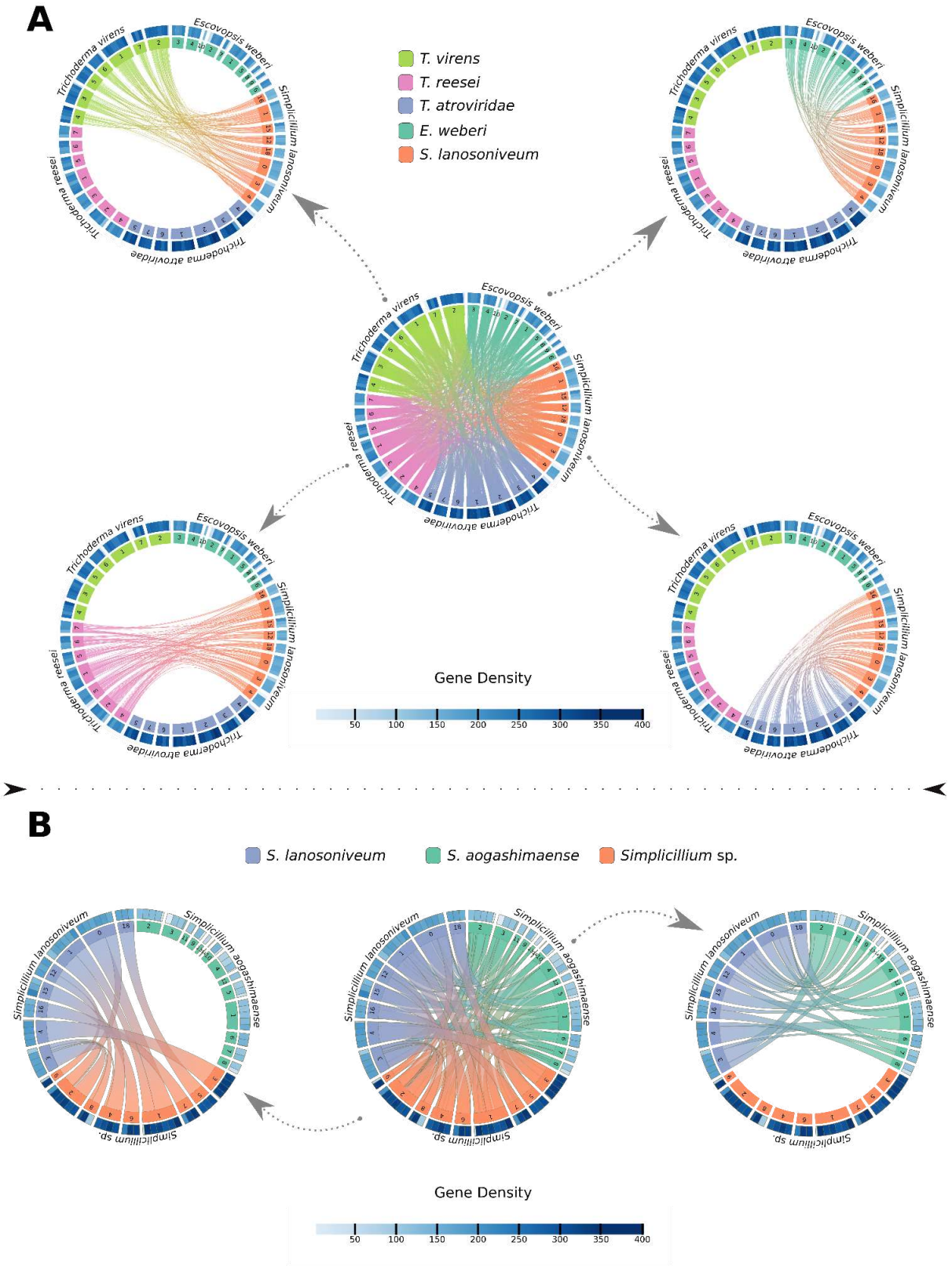



**Figure 4:** Synteny analyses among Hypocreales species (A) and only among *Simplicillium* spp. (B) represented in different colors. The Circo plots represent syntenic blocks that were present in all species analyzed. Rectangles inside the circle represent chromosomes of different species and the lines inside the circle connect syntenic regions of chromosomes from different species. The outer rectangles (heatmap) indicate gene density (the darker blue regions correspond to regions of higher gene density and the lighter blue regions of lower gene density).

**2.4 Cazyme identification and analysis**

The Carbohydrate-active enzymes (Cazymes) are enzymes that are involved in the formation and breakdown of complex carbohydrates and glycoconjugates, including those present in the host tissue. The Cazymes identification was performed with the aim of finding enzymes genus or species-specifics that could be associated with the host specificity. Cazymes are classified into six groups: glycoside hydrolases (GHs), auxiliaries (AAs), carbohydrate-binding modules (CBM), carbohydrate esterases (CEs), polysaccharide lyases (PLs) and glycosyl transferases (GTs). Most of the Cazymes identified (18.35%) were shared with all organisms analyzed, independently to be mycoparasite or not (Figure 5). Seven Cazymes (Glycoside Hydrolase Family (GH) - GH29, GH35, GH43_33, GH84, GH135 Auxiliary Activity Family - AA5_2, GlycosylTransferase Family - GT20) were grouped into two families (Glycosyl Transferase and Glycoside Hydrolase) and shared only among *Simplicillium* spp.. None of Cazymes was shared only among the mycoparasite species (*S. lanosoniveum*, *E. weberi*, *T. virens,* and *T. atroviride*). Four genes were unique to *S. lanosoniveum* and identified as encoding Cazymes (GH30, GH13_31, GH45, GH13_25+GH133) of the Glycoside Hydrolase family. Different Cazymes groups were also evaluated in the *S. lanosoniveum* genome and the most abundant Cazymes observed were GHs, with an average of 64.97%, followed by AAs (13.20%), CBMs (9.64%), CEs (6.09%), GTs (4.57%) and PLs (1.52%) (Figure 5).

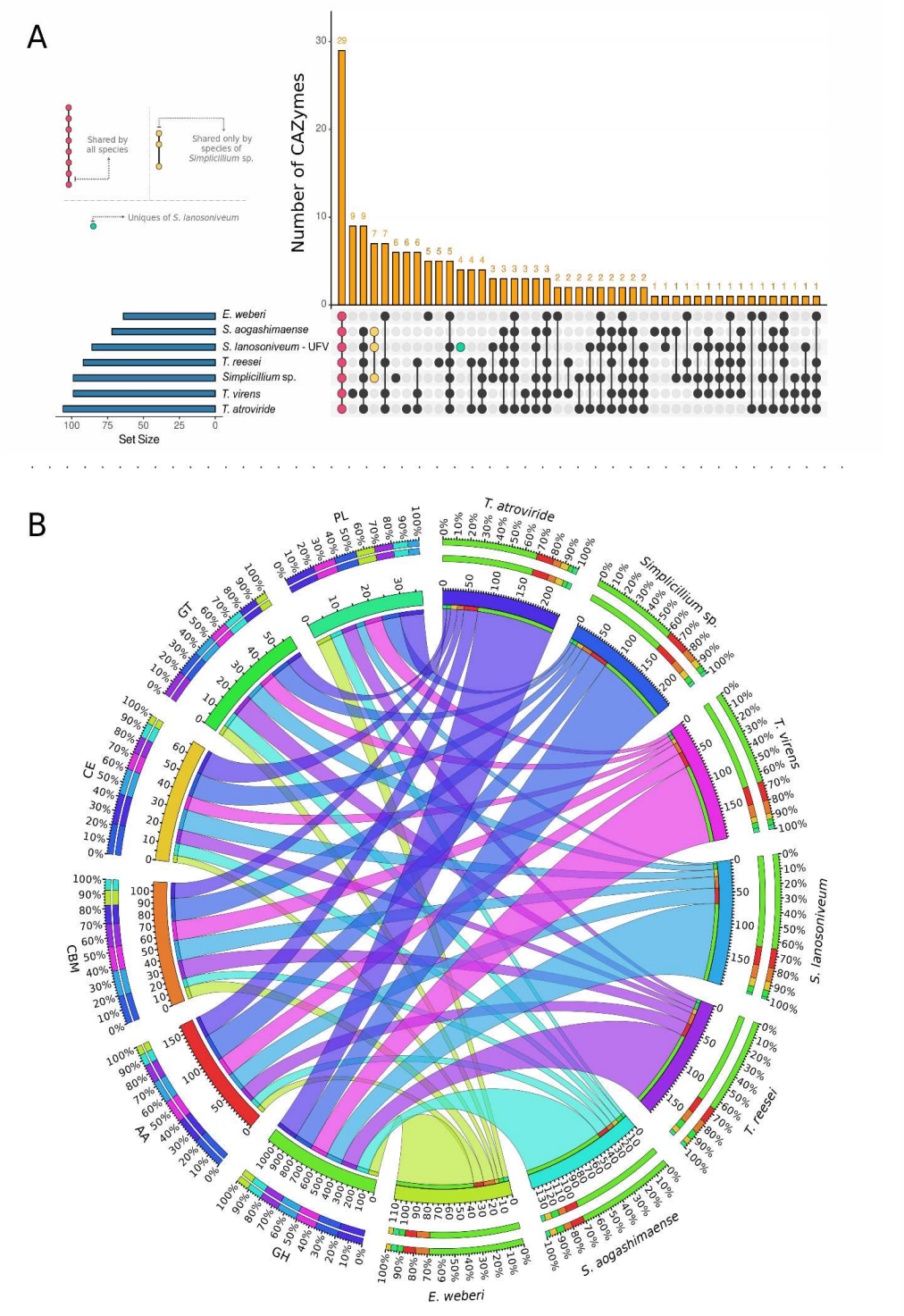



**Figure 5:** Distribution of Carbohydrate-active enzymes (Cazymes) predicted for Hypocreales species (A), lines between dots connect species that share the same Cazymes and the vertical bars over the dots indicate the number of enzymes shared in A. Overview of their relationship (circus plot) with the six Cazymes groups: glycoside hydrolases (GHs), auxiliaries (AAs), carbohydrate-binding modules (CBM), carbohydrate esterases (CEs), polysaccharide lyases (PLs) and glycosyl transferases (GTs) (B). Lines connect the observed number of enzymes annotated from each Cazyme group present per species with the percentage that this amount represents from the total enzymes analyzed for each Cazyme group in B.

**2.5 Secondary metabolism comparison among Hypocreales fungi**

Using AntiSMASH (antibiotics & Secondary Metabolite Analysis Shell), 32 putative gene clusters involved in the biosynthesis of secondary metabolites in *S. lanosoniveum* were identified, including 18 Non-Ribosomal Peptide Synthetase cluster (NRPS) and NRPS-like, eight Type I Polyketide Synthase (T1PKS), five terpenes and one beta-lactone (beta-lactone containing protease inhibitor) also annotated as putative NRPS (Figure 6A). Minimum Information about a Biosynthetic Gene cluster (MIBiG) comparison resulted in variable similarity scores (0,08 - 0,51) against reference Biosynthetic Gene Clusters (BGC). Three clusters predicted to produce Strobilurin, Aculeacin, and Squalestatin metabolites with fungicide activity were selected for completeness analysis. The amino acid sequence of the core gene in the reference BGC was used to explore the number and diversity of the cluster across the genome of different Hypocreales fungi. Furthermore, to improve the investigation of BGC presence in fungi that explore different niches, in addition to the species used in previous analyzes, the genomic sequence from *Cordyceps militaris* was also added to this analysis. *C. militaris* is an entomopathogenic fungus of the order Hypocreales, that also exhibited a high-quality genome assembly available. The number and diversity of the clusters in these species were analyzed by constructing BGC sequence similarity networks, grouping BGCs into gene cluster families, and exploring gene cluster diversity linked to enzyme phylogenies in BIG-SCAPE/CORASON program.

The number of predicted clusters possibly producing Strobilurin and Aculeacin was very different among the analyzed species, varying between 1 to 22 and 4 to 24 cluster copies in each species, respectively (Figure 6B). *T. virens*, *T. atroviride,* and *Simplicillium* sp. were the species with higher copy number of predicted clusters producing Strobilurin and Aculeacin with more than 10 copies per genome of each metabolite cluster, while *S. aogashimaense* has the lowest number of copies, with one and four predicted copies of these metabolites' clusters, respectively. On the other hand, the predicted Squalestatin cluster is the less abundant fungicide cluster, which is present once in the genomes of all analyzed species, except for *Simplicillium* sp. and *S. lanosoniveum* genomes where it was present twice (Figure 6B).

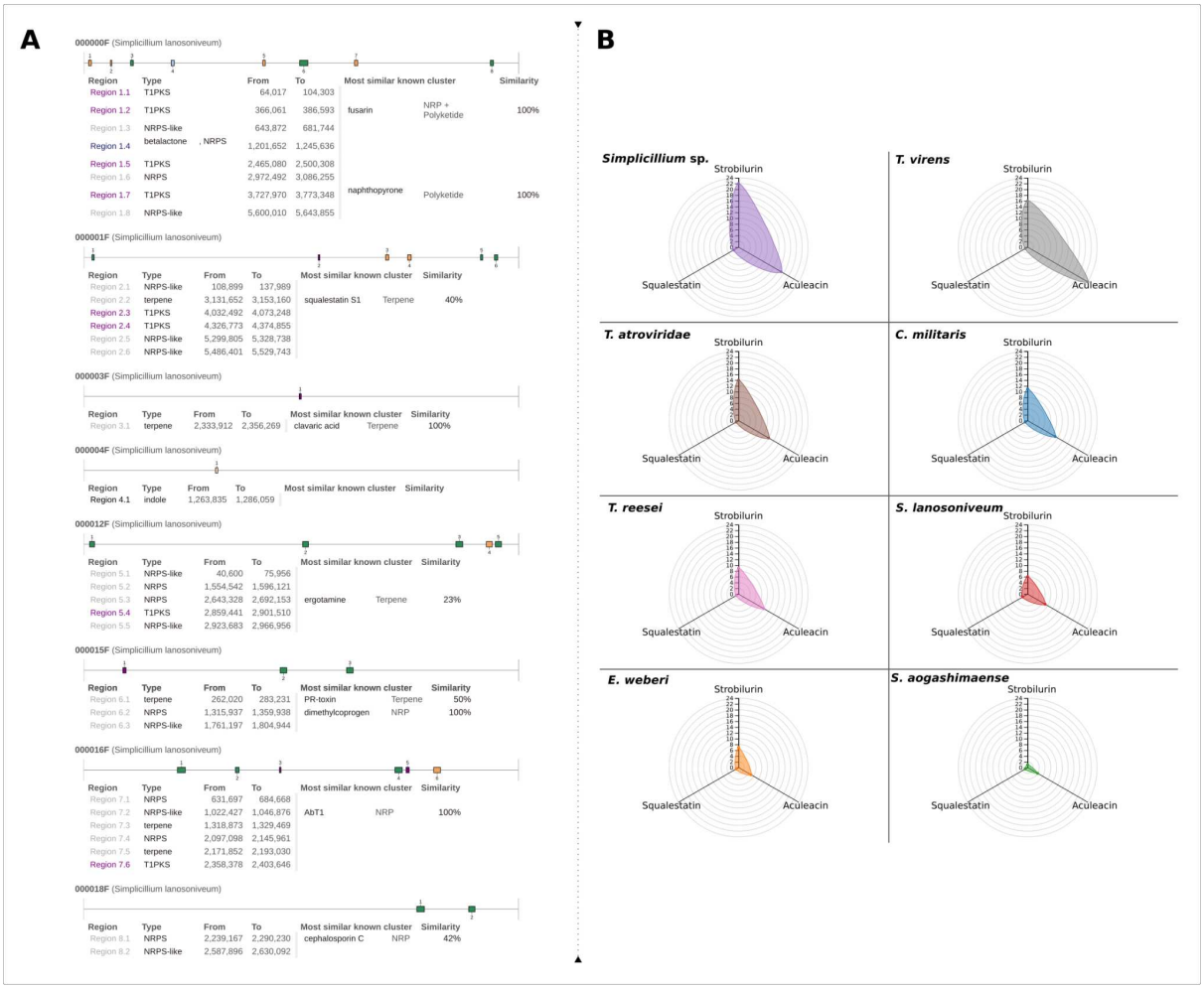



**Figure 6:** Secondary metabolic clusters of SL-UFV isolate identified by AntiSmash. (A) predicted genome position, type of metabolite, and percentage of similarity for conserved clusters. (B) Quantitative comparison of the number of secondary metabolites with fungicide or fungistatic activity (Strobilurin, Acuelacin, and Squalestatin) identified in SL-UFV genome and amount of clusters identified in seven other Hypocreales fungi (*C. militaris, E. weberi*, *Simplicillium* sp., *S. aogashimaense*, *T. atroviride*, *T. reesei,* and *T. virens*) obtained with BigScape.

The similarity network constructed for the Squalestatin cluster indicates that all Hypocreales species analyzed have one similar cluster (SQC1) with one conserved core gene, encoding the Squalestatin synthetase, with lengths between 460 and 472 amino acids (Figure 7). *Simplicillium* spp. form one separated clade with some genus-specific genes (Cluster 1 - Figure 7). Furthermore, *Simplicillium* sp. and *S. lanosoniveum* genomes contain one additional predicted cluster encoding Squalestatin (SQC2) that is unique from these species and very different from that present in all species (Figure 7). SQC2 only has gene homology for the *squalestatin synthetase core gene* and one additional biosynthetic gene (red and light purple boxes, respectively, in Figure 7). Although the conserved function in both clusters, the *squalestatin synthetase core gene* also is shorter in SQC2 (362 amino acids) than the same gene in SQC1, which is found in all the analyzed species. Comparing the predicted

squalestatin synthetase protein structure from both clusters of *S. lanosoniveum* the core part is similar, but the synthetase protein encoded in the cluster restricted to *Simplicillium* sp. and *S. lanosoniveum* is more compact in their barrel structure due to a lack of 38 amino acids (green alpha-helices in protein structures, Figure 7) in their protein sequences at N-terminal extremity, present only in the larger protein (Figure 7). Analysis of the protein domains by Interproscan (Blum et al., 2020) predicted one membrane-bound portion of the protein outside of the membrane, in the cytoplasm.

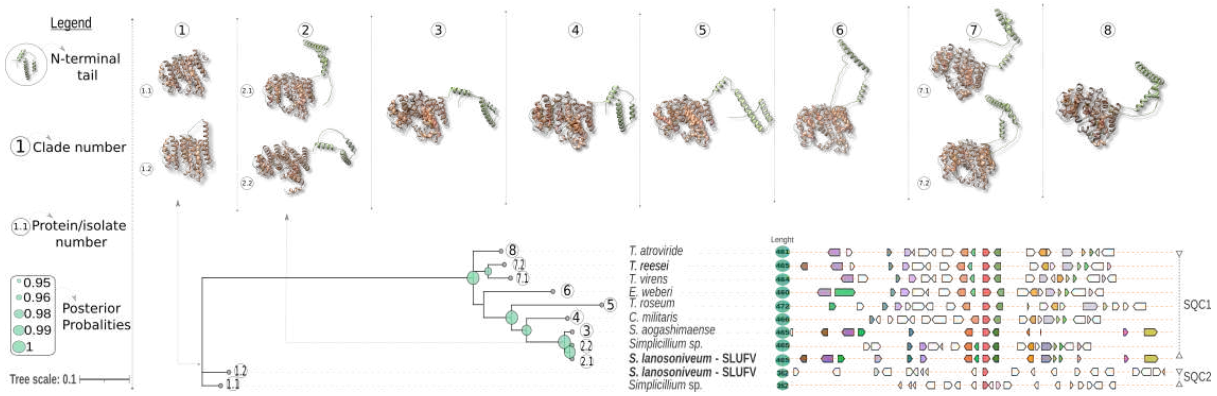



**Figure 7:** Phylogeny of squalestatin clusters SQC1 and SQC2 (of Hypocreales species and the predicted structure of squalestatin synthetase core proteins. Squalestin cluster SQC1 were represented in cluster 2-8 and SQC2 in cluster 1. The tridimensional structure was performed in the Alphafold program and in analysis it was possible to observe the loss of alpha-helices in proteins from SQC2 (Cluster 1) of the *S. lanosoniveum* and *Simplicillium* sp.. Each branch represents one squalestatin cluster with its genes represented in small boxes, that made reference to the size and disposition of the genes in each cluster. Each box was represented by one color, and boxes with the same color represent the same gene. The *squalestatin synthetase core genes* were represented in red and their length was highlighted in the green circles.

## 3. DISCUSSION

A high-quality assembly *S. lanosoniveum* genome with eight contigs was obtained. The presence of telomeric sequences in both extremities of four contigs suggests that four of the eight contigs are complete chromosomes (Figure S1). The genome size and GC content (32.6 Mb, 49.86%) of SL-UFV are closer to *S. aogashimaense* (30,26 Mb, 48,95%). The genome of this related mycoparasitic fungi was assembled in seven pseudo-chromosomes (Zhu et al., 2022), giving additional support to the good quality of the SL-UFV assembled genome. The completeness of the *S. lanosoniveum* genome content was also supported by BUSCO analysis indicating 97.9% of complete genes compared with the Hypocreales database. On another hand, the gene annotation resulted in a smaller number of genes (6779 against 8963) that could be related to different genome evolution or probably due to

differences in genomic tools and parameters used in the annotation process (Han et al., 2013; Hubisz et al., 2011).

Genome organization was compared with other fungi from Hypocreales order using 6682 genes CDS for synteny comparison. The previously reported mesosynteny among genomes of Ascomycetes (Hane et al., 2011) was also observed in our analysis among species from different families of Hypocreales. The genes were conserved within homologous chromosomes, but the order and orientation were randomized along the chromosome, maintaining more similar organization among species in the same genus. The results supported that most of the genome organization is conserved among *Simplicillium* species, implying that specific factors that could be involved with host specialization did not result from massive chromosome rearrangements, but probably are results of the variations in species-specific genes.

Based on studies with other mycoparasite species, mostly *Trichoderma* spp. (Benitez et al., 2004; Lorito et al., 2010; Druzhinina et al., 2012, Karlsson et al., 2017), it is commonly believed that some lytic enzymes such as chitinases, glucanases, and proteases could be determinants for mycoparasitism. Considering that *S. lanosoniveum* needs to secret these enzymes to have their effect on the host, the SignalP tool was used to filter genes predicted to encode secreted proteins. *S. lanosoniveum* secretome contains 969 genes. Among them, twenty-six encodes glucanases, 18 chitinases, and 69 proteases. The number of genes in the secretome is closer to the 968 and 947 genes identified in the *T. atroviride* and *T. virens* secretomes (Druzhinina et al., 2012). The predicted number of proteases also is similar to the 63 proteins identified in *T. virens*, but is lower than the 81 proteases identified in *T. atroviride* (Druzhinina et al., 2012). *S. lanosoniveum* also has fewer predicted chitinase genes (18) than both *T. virens* (36) and *T. atroviride* (29) (Kubicek et al., 2011). The lower number of genes could be associated with the niche of each fungus. *Trichoderma* spp. are species that habit diverse natural and artificial substrata, are adapted to various ecological conditions, and have high opportunistic potential where the antagonism against a diverse community is maintained (Druzhinina and Kubicek, 2013), *S. lanosoniveum* is usually found in soybean pustules from rust pathogens or associated with other rust pathogens where it has one specific host and low competition in its living environment. The high number of secreted enzymes could be related to the success or necessity of each species to survive in their environment.

The niche specialization usually is related to non-essential genes that can be related to substrate metabolisms or antagonism relationships, such as some Cazyme genes or secondary metabolites. Aiming to find the genes shared by all mycoparasitic fungi in their parasitism or

to highlight genes specific from *S. lanosoniveum*, the Cazymes content was compared among eight species from Hypocreales fungi, including the non-mycoparasite species *Simplicillium* sp. and *T. reesei*. This Cazymes analysis showed that most of the genes groups identified (18.35%) are shared with all organisms, while seven Cazymes genes are shared by *Simplicillium* species, and four are unique for *S. lanosoniveum* (GH30, GH13_31, GH45, GH13_25+GH133) belonging to the Glycoside Hydrolase family, a family of enzymes that hydrolyze the bond between a carbohydrate and another compound, such as a second carbohydrate (Tyler et al., 2010). Although four unique Cazymes genes were associated with *S. lanosoniveum,* it was not possible to associate their hydrolase activity with some unique function in mycoparasitism with the information available. Functional experiments must be done to understand if these enzymes could have an undiscovered essential function for *S. lanosoniveum* mycoparasitism.

The predicted secondary metabolite clusters from the *S. lanosoniveum* genome were also compared with other Hypocreales fungi, including mycoparasites, an entomoparasite, a saprophyte, and one *Simplicillium* species isolated from seawater. Looking for secondary metabolites that could be involved in *P. pachyrhizi* parasitism, metabolites with fungicide or fungistatic activity were selected to compare their distribution among the selected fungi species. The number of metabolites with fungicide activity was lower in *S. lanosoniveum* compared with all species of *Trichoderma* spp., indicating that *S. lanosoniveum* could be less competitive against other fungi than *T. reesei*, *T. virens,* and *T. atroviride*, or it could imply that *S. lanosoniveum* is more specialized in parasitism. Differences in the number of toxins produced in *T. reesei*, *T. virens,* and *T. atroviride* species could explain in part the strong competitiveness and success of *Trichoderma* spp. in nature and their use as generalist agents of biological control against a broad variety of plant pathogens (Zin and Badaluddin, 2020), while *S. lanosoniveum* has been less reported in nature, sometimes restricted to an antagonistic relationship with aerial plant pathogens, exhibiting a restricted use as an agent of biological control. This specificity can be a great advantage if used to control soybean rust since the effects will be specific to the pathogen with a small effect on other organisms beneficial to soybean development.

Despite *Trichoderma* spp., in general, having more clusters of secondary metabolites with fungicide activity than *S. lanosoniveum*, the cluster associated with Squalestatin production was observed to be an exception. Squalestatin is usually produced by several ascomycetes and exhibits broad antifungal activity, acting as a potent inhibitor of squalene synthase, an essential enzyme for sterol biosynthesis, a lipid component of the cell membrane

(Bergstrom et al., 1995). In the genomic analysis of *S. lanosoniveum*, one Squalestatin cluster (SQC1) containing a similar amount of homolog genes and similar structures in all analyzed species was found. This observation agrees with the previous information about the vast distribution of this cluster in Ascomycetes (Bills et al., 1994). In addition, a second predicted Squalestatin cluster (SQC2) exclusive of *S. lanosoniveum* and *Simplicillium* sp. was also identified. SQC2 differs in the number and function of most of the genes and has only the *squalestatin synthetase core gene* and one additional biosynthetic gene that are from gene families also presented in SQC1. Although the possible conserved function, *squalestatin synthetase core genes* from SQC1 and SQC2 exhibited differences in size, protein conformation, and domains. These differences in the core gene and differences in the cluster could result in different squalestatin products with distinct targets or alternative destinies inside or outside of the cell. More than one kind of Squalestatin is known and probably the second cluster present in *S. lanosoniveum*, and absent in other fungi including *S. aogashimaense* (not a rust fungi parasite), encodes a different compound that could be necessary and specific to the parasitism of rust pathogens, such as *P. pachyrhizi.* The Squalestatin protein is reported to be linked to the membranes; however, the absence of one transmembrane domain, which would bind the protein to the membrane in the *squalestatin synthetase core gene* of SQC2 was observed. This alteration could modify the protein location, and therefore the function. To confirm this hypothesis and the essentiality of this toxin for *Simplicillium* mycoparasites species against *P. pachyrhizi*, further functional studies will be necessary including subcellular localization, gene silencing experiments, and obtaining knockout mutants for essential genes of SQC2.

## 4. CONCLUSION

In this study, the first reference genome assembly with functional annotation of the mycoparasitic fungus *S. lanosoniveum* was obtained. Conserved genome organization (macrosynteny) among *Simplicillium* spp. and the mesosynteny among Hypocreales fungi was found. Genomic features shared between different mycoparasitic species, and some unique genes and clusters associated with *Simplicillium* species were also identified. A possible Squalestatin cluster present in the rust parasite species *S. lanosoniveum* and absent in other mycoparasite species may be associated with *P. pachyrhizi* infection. The function confirmation of candidate parasitic genes and clusters associated with *P. pachyrhizi*

mycoparasitism will require further studies such as gene silencing experiments and development of knockout mutants, among others.

## 5. MATERIALS AND METHODS

### 5.1 Mycoparasite isolation

The isolate SL-UFV was isolated from uredinias of *P. pachyrhizi* in soybean leaves inoculated and maintained at a growth chamber for experiments in Viçosa (Minas Gerais state, Brazil). Conidia from the mycoparasite fungus on the uredinia were collected and transferred to Potato dextrose agar (PDA) plates and incubated for two weeks at 25°C. SL-UFV was routinely maintained on PDA (potato dextrose agar) slant and stored at −80 °C as glycerol stock.

To reproduce the parasitism of SL-UFV on *P. pachyrhizi,* the isolate was grown on aseptic PDA medium for two weeks and the colony was washed with distilled water to obtain a conidia suspension. The suspension was filtered with cheesecloth and adjusted to 9 x $10^4$ conidia $mL^{-1}$. Inoculations were performed on detached soybean leaves previously inoculated with *P. pachyrhizi* 14 days before the *S. lanosoniveum* inoculation. The control treatment was sprayed with distilled water. All the leaves inoculated were maintained in Petri dishes with wet cotton for two weeks at room temperature (approximately 23°C) and photos (Figure 1) were taken in the stereo microscope.

### 5.2 Genome and RNA sequencing

SL-UFV mycelia and spores harvested from four-day-old PDB (potato dextrose broth) culture in 150 rpm shaker at 25 °C were used to isolate high molecular weight DNA using CTAB (cetyltrimethylammonium bromide) method (Jones & Schewessinger, 2021). A total of 40 µg purified genomic DNA was used to construct a standard PacBio SMRTbell library using PacBio SMRT Express Template Prep Kit 2.0 (Pacific Biosciences, CA). The sequencing was performed using a PacBio Sequel II instrument at BGI (https://www.bgi.com) with 15 Kb a Blue Pippin size selection being performed prior to sequencing. Sequence data collection was standardized to 30 hours to allow ample time for multiple pass sequencing around SMRTbell template molecules of 10–25 Kb which yields high-quality circular consensus sequencing (HiFi) results. Raw base-called data was moved from the sequencing instrument and imported into SMRTLink to generate HiFi reads using the CCS algorithm

which processed the raw data and generated the HiFi fastq files with the following settings: minimum pass 3, minimum predicted RQ 20.

Detached soybean leaflets were inoculated with 9 x $10^4$ conidia $mL^{-1}$ of SL-UFV isolate, 14 days after inoculation with *P. pachyrhizi*. Total RNA used in the annotation process was isolated from leaflets samples collected 0, 12, 24, 36, 48, 96 e 120 hours after SL-UFV inoculation using the protocol described by Bilgin et al. (2009). Illumina RNA-seq libraries were prepared from total purified polyA RNA and sequenced-by-synthesis with TruSeq v3 chemistry on a HiSeq2000 at Novogene (USA). Reads from *Glycine max* transcriptome were filtered out using fastq_screen v0.4.1 (www.bioinformatics.babraham.ac.uk/projects/fastq_screen) as described by Kellner *et al.* (2014) and used in the genome annotation pipeline.

**5.3 Sequencing, genome assembly and annotation**

HiFi reads were assembled using hifiasm (Cheng et al., 2021) with default parameters. Tapestry (https://github.com/johnomics/tapestry) was used to search for telomeric regions in all contigs using the default configuration.

Gene prediction and annotation were performed using the Funannotate pipeline in pattern conditions (Palmer, 2017). The complete pipeline was also able to remove small repetitive contigs with the minimap2 program (Li, 2018) and mummer algorithms (Kurtz et al., 2004). The minimum number of training models required was 200, and the Optimize Augustus was turned on. Funannotate also uses the programs AUGUSTUS, and tRNAscan-SE to predict genes for proteins and tRNAs. To increase predictions, Funannotate was set up to run the predictors Phobius (Käll et al., 2004) and AntiSMASH (Blin et al., 2021). Finally, the annotation function of the Funannotate was used to create the final genome annotation file. BUSCO program (Simão et al., 2015) was used to quantify the gene completeness, using the same database of the assembly phase (Hypocreales_odb10).

Funannotate was also executed to identify numbers and classes of carbohydrate-active enzymes (Cazymes) and to predict the secreted proteins (using SignalP5) just in one software. CAZymes were reannotated using the dbCAN2 meta server (Zhang et al., 2018). Genome assembly results and annotation were submitted to BioKit (Steenwyk et al., 2022) for reporting some information about genome assembly quality and summary statistics. The package Circlize implemented in R software was used to create Figure S1 and the Tapestry tool was used to identify telomeric sequences (https://github.com/johnomics/tapestry).

### 5.4 Phylogenetic analyses

For phylogenetic analyses, regions used to identify *Simplicillium* sp. (ITS, LSU, and SSU) (Chen et al., 2021; Wei et al., 2019) were filtered and used. The database to construct phylogeny was done with sequences deposited in Genbank (Clark et al., 2016) and described in Supplementary Table 1.

The sequences were aligned by the MAFFT v.7 (Katoh et al., 2019) with default options, and the alignment curation was performed using GBlocks v.0.91b (Talavera and Castresana, 2007). A partition homogeneity test (PHT) between the SSU, ITS , and LSU sequences was performed with PAUP 4.0a (Cummings, 2014).

The evolutionary story of these genes from *Simplicillium* was reconstructed, using Bayesian inference (BI) analyses, based on the Markov Chain Monte Carlo (mcmc) method, to confirm the identity of the isolate. The best nucleotide substitution model was defined by running MrMODELTEST 2.3 (Posada and Buckley, 2004). Based on the Akaike Information Criterion (AIC), the selected models were SYM+I+G (ITS), GTR+G (LSU), and F81 (SSU). The BI analysis was performed in MrBayes v.3.1.1 (Huelsenbeck et al., 2001) using four MCMC chains simultaneously with the following setup: ngen = 20000000, samplefreq = 5000 and burnin = 10%. We analyzed the likelihood log convergence in TRACER v. 1.4.1 (Rambaut et al., 2018) and visualized and edited the tree using iTOL v.5 (Letunic and Bork, 2021). Tree was rooted in *Pochonia chlamydosporia* CBS103.65 isolate.

### 5.5 Analysis of genome-wide synteny

Six other genomes from Hypocreales fungi with good assembly and annotation quality were also used for synteny analysis: two *Simplicillium* species [*S. aogashimaense* (GCA_012273805.1) and *Simplicillium* sp. (GCA_022702485.1)], three fungi species also known as mycoparasite (*T. atroviride* (GCA_020647795.1)*, T. virens* (GCA_020647635.1) and *E. weberi* (GCA_003055145.1)] and the non-mycoparasite fungus *T. reesei* (GCA_002006585.1).

To perform the comparative genome analysis, orthologous gene pairs were identified using the annotation results reported by Funannotate. The alignments were performed using BLASTP with an e-value < $1e^{-10}$. The collinearity analyses were conducted using the MCScan X (Wang et al., 2012), and the gene density for each genome was calculated using the RIdeigram package (Hao et al., 2020) on R software (R Core Team, 2020). The results of collinearity and the gffs files were used with the ACCUSYN tool for building and visualizing

the syntenic blocks and their relationships, as well as the gene density of each analyzed species.

### 5.6 Secondary metabolism comparison using genome mining

The identification of the secondary metabolism clusters was done using the antiSMASH 6.0 fungi version web server (https://fungismash.secondarymetabolites.org/#!/start) using detection strictness: relaxed and all extra features on. The same six species described in the previous topic and additional data of the species *Cordyceps militaris* (GCA_008080495.1), an entomopathogenic fungus of the order Hypocreales were included to improve the comparison with species from different niches. The secondary metabolite clusters using the reference Biosynthetic Gene Cluster (BCG) were annotated with the highest similarity score on the MIBiG comparison results from antiSMASH.

Secondary metabolites cluster predictions obtained in AntiSMASH were submitted to BiG-SCAPE (Biosynthetic Gene Similarity Clustering and Prospecting Engine) and CORASON (CORe Analysis of Syntenic Orthologs to prioritize Natural Product Biosynthetic Gene Clusters) (Navarro-Muñoz et al., 2020). Using genome mining approaches to explore the diversity of BGCs and constructing BGC sequence similarity networks, Big-SCAPE and CORASON were used to construct enzyme phylogenies. For squalestatin clusters, the predicted protein encoded by the *squalestatin synthetase core gene* from SQC1 and SQC2 were filtered and the three-dimensional structures were predicted from the sequences using a combination of MMseqs2 and AlphaFold2, according to the recommendations in their manuals (Jumper et al., 2021; Mirdita et al., 2022). Identification of domains and comparison of the predicted proteins encoded by *squalestatin synthetase core gene* from SQC1 and SQC2 were performed in the Interproscan web server (https://www.ebi.ac.uk/interpro/).

## 6. REFERENCES


Atanasova, L., Crom, S. L., Gruber, S., Coulpier, F., Seidl-Seiboth, V., Kubicek, C. P., & Druzhinina, I. S. (2013). Comparative transcriptomics reveals different strategies of *Trichoderma* mycoparasitism. *BMC Genomics*, 14, 1-15.

Barnett, H. L. (1963). The nature of mycoparasitism by fungi. *Annual Reviews in Microbiology*, 17 (1), 1-14.

Benítez, T., Rincón, A. M., Limón, M. C., & Codon, A. C. (2004). Biocontrol mechanisms of *Trichoderma* strains. *International Microbiology*, 7 (4), 249-260.

Bergstrom, J. D., Dufresne, C., Bills, G. F., Nallin-Omstead, M., & Byrne, K. (1995). Discovery, biosynthesis, and mechanism of action of the zaragozic acids: potent inhibitors of squalene synthase. *Annual review of microbiology*, 49 (1), 607-639.

Bilgin, D. D., DeLucia, E. H., & Clough, S. J. (2009). A robust plant RNA isolation method suitable for Affymetrix GeneChip analysis and quantitative real-time RT-PCR. *Nature protocols*, 4 (3), 333-340.

Bills, G. F., Peláez, F., Polishook, J. D., Diez-Matas, M. T., Harris, G. H., Clapp, W. H., ... & Bergstrom, J. D. (1994). Distribution of zaragozic acids (squalestatins) among filamentous ascomycetes. *Mycological research*, 98 (7), 733-739.

Blin, K., Shaw, S., Kloosterman, A. M., Charlop-Powers, Z., Van Wezel, G. P., Medema, M. H., & Weber, T. (2021). antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic acids research*, 49 (W1), 29-35.

Blum, M., Chang, H. Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., ... & Finn, R. D. (2021). The InterPro protein families and domains database: 20 years on. *Nucleic acids research*, 49 (D1), 344-354.

Cai, F., & Druzhinina, I. S. (2021). In honor of John Bissett: authoritative guidelines on molecular identification of *Trichoderma*. *Fungal Diversity*, 107, 1-69.

Chen, W. H., Han, Y. F., Liang, J. D., & Liang, Z. Q. (2021). Taxonomic and phylogenetic characterizations reveal four new species of *Simplicillium* (Cordycipitaceae, Hypocreales) from Guizhou, China. *Scientific Reports*, 11 (1), 15300.

Clark, K., Karsch-Mizrachi, I., & Lipman, D. J. (2016). GenBank. *Nucleic Acids Research*, 44, 67–72.

Cummings, M. P. (2004). PAUP (phylogenetic analysis using parsimony (and other methods)). *Dictionary of Bioinformatics and Computational Biology*.

de Man, T. J., Stajich, J. E., Kubicek, C. P., Teiling, C., Chenthamara, K., Atanasova, L., ... & Gerardo, N. M. (2016). Small genome of the fungus *Escovopsis weberi*, a specialized disease agent of ant agriculture. *Proceedings of the National Academy of Sciences*, 113 (13), 3567-3572.

Druzhinina, I. S., Shelest, E., & Kubicek, C. P. (2012). Novel traits of *Trichoderma* predicted through the analysis of its secretome. *FEMS microbiology letters*, 337 (1), 1-9.

Druzhinina, I. S., & Kubicek, C. P. (2013). Ecological genomics of *Trichoderma*. *The Ecological Genomics of Fungi*, 89-116.

Fukuda, T., Sudoh, Y., Tsuchiya, Y., Okuda, T., & Igarashi, Y. (2014). Isolation and biosynthesis of preussin B, a pyrrolidine alkaloid from *Simplicillium lanosoniveum*. *Journal of Natural Products*, 77 (4), 813-817.

Gauthier, N. W., Maruthachalam, K., Subbarao, K. V., Brown, M., Xiao, Y., Robertson, C. L., & Schneider, R. W. (2014). Mycoparasitism of *Phakopsora pachyrhizi*, the soybean rust pathogen, by *Simplicillium lanosoniveum*. *Biological Control*, 76, 87-94.

Han, M. V., Thomas, G. W., Lugo-Martinez, J., & Hahn, M. W. (2013). Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Molecular biology and evolution*, 30 (8), 1987-1997.

Hane, J. K., Rouxel, T., Howlett, B. J., Kema, G. H., Goodwin, S. B., & Oliver, R. P. (2011). A novel mode of chromosomal evolution peculiar to filamentous Ascomycete fungi. *Genome biology*, 12, 1-16.

Hao, Z., Lv, D., Ge, Y., Shi, J., Weijers, D., Yu, G., & Chen, J. (2020). RIdeogram: drawing SVG graphics to visualize and map genome-wide data on the idiograms. *PeerJ Computer Science*, 6, e251.

Heine, D., Holmes, N. A., Worsley, S. F., Santos, A. C. A., Innocent, T. M., Scherlach, K., ... & Wilkinson, B. (2018). Chemical warfare between leafcutter ant symbionts and a co-evolved pathogen. *Nature Communications*, 9 (1), 2208.

Hubisz, M. J., Lin, M. F., Kellis, M., & Siepel, A. (2011). Error and error mitigation in low-coverage genome assemblies. *PloS one*, 6 (2), e17034..

Huelsenbeck, J. P., Ronquist, F., Nielsen, R., & Bollback, J. P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, 294 (5550), 2310-2314.

Jones, A., Nagar, R., Sharp, A., & Schwessinger, B. (2019). High-molecular weight DNA extraction from challenging fungi using CTAB and gel purification. *Protocols. io online*.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596 (7873), 583-589.

Käll, L., Krogh, A., & Sonnhammer, E. L. (2004). A combined transmembrane topology and signal peptide prediction method. *Journal of molecular biology*, 338 (5), 1027-1036.

Karlsson, M., Atanasova, L., Jensen, D. F., & Zeilinger, S. (2017). Necrotrophic mycoparasites and their genomes. *Microbiology Spectrum*, 5 (2), 5-2.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30 (4), 772-780.

Kellner, R., Bhattacharyya, A., Poppe, S., Hsu, T. Y., Brem, R. B., & Stukenbrock, E. H. (2014). Expression profiling of the wheat pathogen *Zymoseptoria tritici* reveals genomic patterns of transcription and host-specific regulatory programs. *Genome biology and evolution*, 6 (6), 1353-1365.

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome research*, 27 (5), 722-736.

Kosawang, C., Karlsson, M., Vélëz, H., Rasmussen, P. H., Collinge, D. B., Jensen, B., & Jensen, D. F. (2014). Zearalenone detoxification by zearalenone hydrolase is important for the antagonistic ability of *Clonostachys rosea* against mycotoxigenic *Fusarium graminearum*. *Fungal Biology*, 118 (4), 364-373.

Krasnov, G. S., Pushkova, E. N., Novakovskiy, R. O., Kudryavtseva, L. P., Rozhmina, T. A., Dvorianinova, E. M., ... & Melnikova, N. V. (2020). High-quality genome assembly of *Fusarium oxysporum* f. sp. *lini*. *Frontiers in genetics*, 11, 959.

Kubicek, C. P., Herrera-Estrella, A., Seidl-Seiboth, V., Martinez, D. A., Druzhinina, I. S., Thon, M., ... & Grigoriev, I. V. (2011). Comparative genome sequence analysis underscores mycoparasitism as the ancestral life style of *Trichoderma*. *Genome biology*, 12, 1-15.

Kubicek, C. P., Steindorff, A. S., Chenthamara, K., Manganiello, G., Henrissat, B., Zhang, J., ... & Druzhinina, I. S. (2019). Evolution and comparative genomics of the most common *Trichoderma* species. *BMC genomics*, 20, 1-24..

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., & Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome biology*, 5, 1-9.

Letunic, I., & Bork, P. (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, 23 (1), 127-128.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, *34* (18), 3094-3100.

Lorito, M., Woo, S. L., Harman, G. E., & Monte, E. (2010). Translational research on *Trichoderma*: from 'omics to the field. *Annual review of phytopathology*, 48, 395-417.

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., & Steinegger, M. (2022). ColabFold: making protein folding accessible to all. *Nature methods*, 19 (6), 679-682.

Navarro-Muñoz, J. C., Selem-Mojica, N., Mullowney, M. W., Kautsar, S. A., Tryon, J. H., Parkinson, E. I., ... & Medema, M. H. (2020). A computational framework to explore large-scale biosynthetic diversity. *Nature chemical biology*, 16 (1), 60-68.

Nygren, K., Dubey, M., Zapparata, A., Iqbal, M., Tzelepis, G. D., Durling, M. B., ... & Karlsson, M. (2018). The mycoparasitic fungus *Clonostachys rosea* responds with both common and specific gene expression during interspecific interactions with fungal prey. *Evolutionary applications*, 11 (6), 931-949.

Palmer, J., & Stajich, J. E. (2017). Funannotate: eukaryotic genome annotation pipeline. *Zenodo*. doi, 10.

Posada, D., & Buckley, T. R. (2004). Advantages of AIC and Bayesian approaches over likelihood ratio tests for model selection in phylogenetics. *Systematic Biology*, 53, 793-808.

R Core Team. (2020). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. *http://www. R-project. org/*.

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., & Suchard, M. A. (2018). Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Systematic biology*, 67 (5), 901-904.

Rukachaisirikul, V., Chinpha, S., Saetang, P., Phongpaichit, S., Jungsuttiwong, S., Hadsadee, S., ... & Ingkaninan, K. (2019). Depsidones and a dihydroxanthenone from the endophytic fungi *Simplicillium lanosoniveum* (JFH Beyma) Zare & W. Gams PSU-H168 and PSU-H261. *Fitoterapia*, 138, 104286.

Shin, T. S., Yu, N. H., Lee, J., Choi, G. J., Kim, J. C., & Shin, C. S. (2017). Development of a biofungicide using a mycoparasitic fungus *Simplicillium lamellicola* BCP and its control efficacy against gray mold diseases of tomato and ginseng. *The plant pathology journal*, 33 (3), 337.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31 (19), 3210-3212.

Spatafora, J. W., Sung, G. H., Sung, J. M., Hywel‐Jones, N. L., & White Jr, J. F. (2007). Phylogenetic evidence for an animal pathogen origin of ergot and the grass endophytes. *Molecular ecology*, 16 (8), 1701-1711.

Steenwyk, J. L., Buida III, T. J., Gonçalves, C., Goltz, D. C., Morales, G., Mead, M. E., ... & Rokas, A. (2022). BioKIT: a versatile toolkit for processing and analyzing diverse types of sequence data. *Genetics*, 221 (3), iyac079.

Talavera, G., & Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic biology*, 56 (4), 564-577.

Tyler, L., Bragg, J. N., Wu, J., Yang, X., Tuskan, G. A., & Vogel, J. P. (2010). Annotation and comparative analysis of the glycoside hydrolase genes in *Brachypodium distachyon*. *BMC genomics*, 11 (1), 1-21.

Wang, N., Fan, X., Zhang, S., Liu, B., He, M., Chen, X., ... & Wang, X. (2020). Identification of a Hyperparasitic *Simplicillium obclavatum* Strain Affecting the Infection Dynamics of *Puccinia striiformis* f. sp. *tritici* on Wheat. *Frontiers in Microbiology*, 11, 1277.

Wang, Y., Tang, H., Debarry, J. D., Tan, X., Li, J., Wang, X., ... & Paterson, A. H. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic acids research*, 40 (7), e49-e49.

Ward, N. A., Schneider, R. W., & Aime, M. C. (2011). Colonization of soybean rust sori by *Simplicillium lanosoniveum*. *Fungal Ecology*, 4 (5), 303-308.

Ward, N. A., Robertson, C. L., Chanda, A. K., & Schneider, R. W. (2012). Effects of *Simplicillium lanosoniveum* on *Phakopsora pachyrhizi*, the soybean rust pathogen, and its use as a biological control agent. *Phytopathology*, 102 (8), 749-760.

Wei, D. P., Wanasinghe, D. N., Hyde, K. D., Mortimer, P. E., Xu, J., Xiao, Y. P., ... & To-Anun, C. (2019). The genus *Simplicillium*. *MycoKeys*, 60, 69.

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P. C., Hall, R. J., Concepcion, G. T., ... & Hunkapiller, M. W. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, 37 (10), 1155-1162.

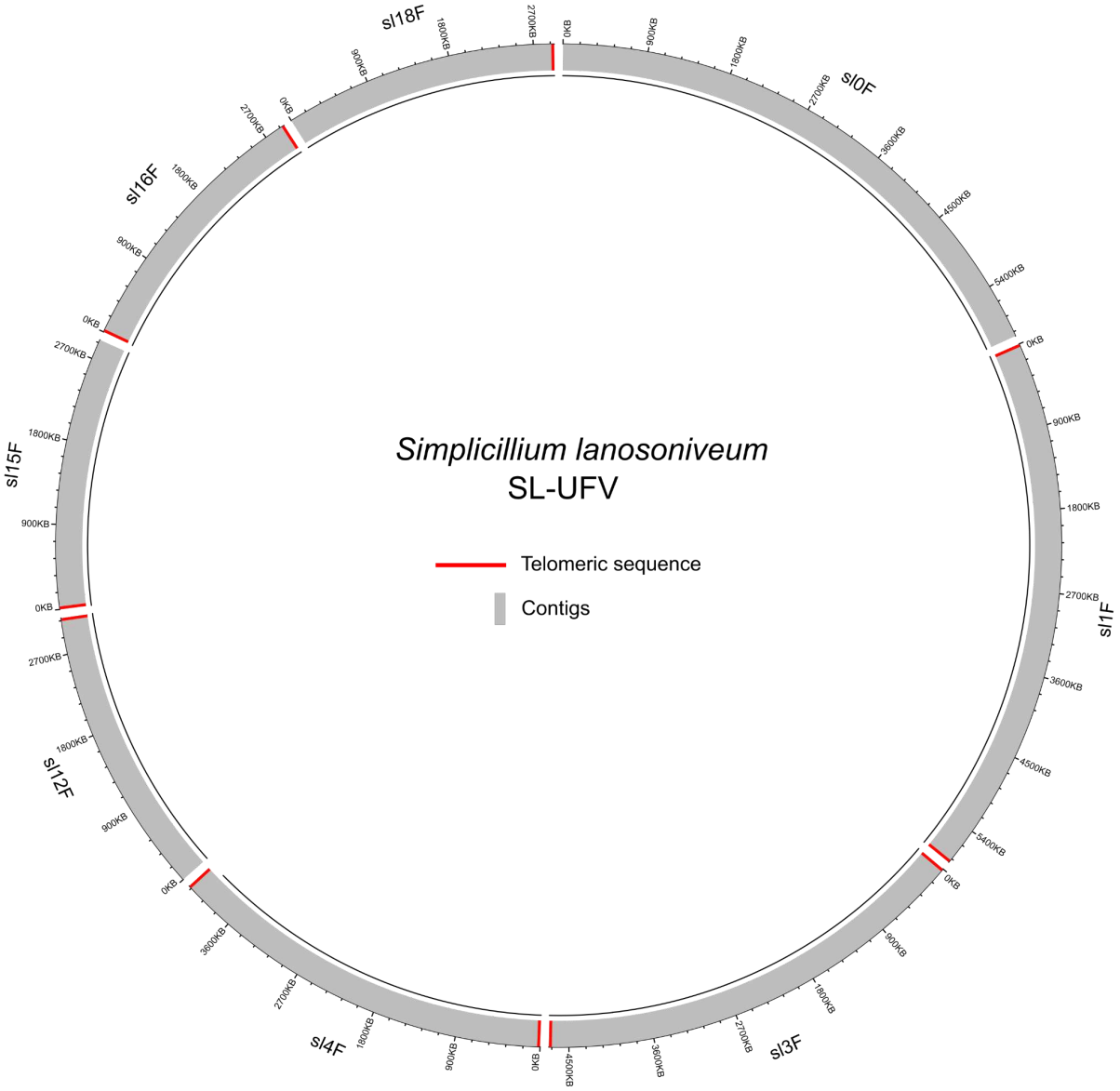
Wilson, A., Cuddy, W. S., Park, R. F., Harm, G. F. S., Priest, M. J., Bailey, J., & Moffitt, M. C. (2020). Investigating hyperparasites as potential biological control agents of rust pathogens on cereal crops. *Australasian Plant Pathology*, 49, 231-238.

Zhang, H., Yohe, T., Huang, L., Entwistle, S., Wu, P., Yang, Z., ... & Yin, Y. (2018). dbCAN2: A meta server for automated carbohydrate-active enzyme annotation. *Nucleic*, 1343, 95-101.

Zhu, M., Duan, X., Cai, P., Li, Y. F., & Qiu, Z. (2022). Deciphering the genome of *Simplicillium aogashimaense* to understand its mechanisms against the wheat powdery mildew fungus *Blumeria graminis* f. sp. *tritici*. *Phytopathology Research*, 4 (1), 16.

Zin, N. A., & Badaluddin, N. A. (2020). Biological functions of *Trichoderma* spp. for agriculture applications. *Annals of Agricultural Sciences*, 65 (2), 168-178.

## SUPPLEMENTARY MATERIAL




**Supplementary Figure 1**: Representation of SL-UFV assembled nuclear genome. The grey bars represent the eight contigs and their size and the red bars in the extremity of the contigs represent telomeric repeats identified with Tapestry.

**Supplementary Table 1:** Genbank accessions used to make phylogenetic reconstructions.

| Species | Strain no. | GenBank accession no. | | |
|---|---|---|---|---|
| | | ITS | SSU | LSU |
| *Simplicillium aogashimaense* | JCM 18167 | AB604002 | | |
| *S. aogashimaense* | JCM 18168 | AB604004 | | |
| *S. calcicola* | LC5371 | KU746705 | | KU74675 |
| *S. calcicola* | LC5586 | KU746706 | | KU746752 |
| *S. chinense* | LC1342 | JQ410323 | | JQ410321 |
| *S. chinense* | LC1345 | NR 155782 | | JQ410322 |
| *S. cicadellidae* | GY11011 | MN006243 | | |
| *S. cicadellidae* | GY11012 | MN006244 | | |
| *S. coffeanum* | COAD 2057 | MF066034 | | MF066032 |
| *S. coffeanum* | COAD 2061 | MF066035 | | MF066033 |
| *S. cylindrosporum* | JCM 18169 | AB603989 | | |
| *S. cylindrosporum* | JCM 18170 | AB603994 | | |
| *S. cylindrosporum* | JCM 18171 | AB603997 | | |
| *S. cylindrosporum* | JCM 18172 | AB603998 | | |
| *S. cylindrosporum* | JCM 18173 | AB603999 | | |
| *S. cylindrosporum* | JCM 18174 | AB604005 | | |
| *S. cylindrosporum* | JCM 18175 | AB604006 | | |
| *S. filiforme* | URM 7918 | MH979338 | | MH979399 |
| *S. formicae* | MFLUCC 18–1379 | MK766511 | MK765046 | MK766512 |
| *S. formicidae* | DL10041 | MN006241 | | |
| *S. formicidae* | DL10042 | MN006242 | | |
| *S. lamellicola* | CBS 116.25 | AJ292393 | | AF339552 |
| *S. lamellicola* | KYK00006 | AB378533 | | |
| *S. lamellicola* | UAMH 2055 | AF108471 | | |
| *S. lamellicola* | UAMH 4785 | AF108480 | | |

| | | | | |
|---|---|---|---|---|
| *S. lanosoniveum* | CBS 101267 | AJ292395 | | AF339553 |
| *S. lanosoniveum* | CBS 704.86 | AJ292396 | | AF339554 |
| *S. lepidopterorum* | GY29131 | MN006246 | | |
| *S. lepidopterorum* | GY29132 | MN006245 | | |
| *S. minatense* | JCM 18176 | AB603992 | LC496893 | |
| *S. minatense* | JCM 18177 | AB603991 | | |
| *S. minatense* | JCM 18178 | AB603993 | LC496894 | |
| *S. obclavatum* | CBS 311.74 | AJ292394 | | AF339517 |
| *S. obclavatum* | JCM 18179 | AB604000 | | |
| *S. spumae* | JCM 39050 | LC496869 | LC496898 | LC496883 |
| *S. spumae* | JCM 39051 | LC496870 | LC496899 | LC496884 |
| *S. spumae* | JCM 39054 | LC496871 | LC496902 | LC496887 |
| *S. subtropicum* | JCM 18180 | AB603990 | LC496895 | |
| *S. subtropicum* | JCM 18181 | AB603995 | LC496896 | |
| *S. subtropicum* | JCM 18182 | AB603996 | | |
| *S. subtropicum* | JCM 18183 | AB604001 | | |
| *S. sympodiophorum* | JCM 18184 | AB604003 | LC496897 | |
| *S. coccinellidae* | DY101791 | MT453861 | MT453863 | MT453862 |
| *S. coccinellidae* | DY101792 | MT453864 | | MT457410 |
| *S. hymenopterorum* | DY101691 | MT453848 | MT453849 | MT453850 |
| *S. hymenopterorum* | DY101692 | MT453851 | MT453852 | MT453853 |
| *S. neolepidopterorum* | DY101751 | MT453854 | MT453856 | MT453855 |
| *S. neolepidopterorum* | DY101752 | MT453857 | MT453859 | MT453858 |
| *S. scarabaeoidea* | DY101391 | MT453842 | MT453843 | MT453844 |
| *S. scarabaeoidea* | DY101392 | MT453845 | MT453847 | MT453846 |
| *Pochonia chlamydosporia* | CBS 103.65 | MH858504 | | |