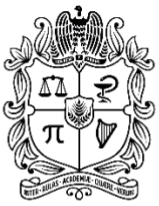


UNIVERSIDAD
NACIONAL
DE COLOMBIA

High-quality genome assembly and comparative genomics of *Pseudocercospora ulei* GCL012, the causal agent of the South American leaf blight (SALB) in natural rubber tree *Hevea brasiliensis*: Towards the prediction of molecular components associated with its pathogenicity and virulence

Sandra Milena González Sayer

Universidad Nacional de Colombia

Sciences Faculty, Biotechnology PhD program

Bogotá, Colombia

2021

High-quality genome assembly and comparative genomics of *Pseudocercospora ulei* GCL012, the causal agent of the South American leaf blight (SALB) in natural rubber tree *Hevea brasiliensis*: Towards the prediction of molecular components associated with its pathogenicity and virulence.

Sandra Milena Gonzalez Sayer

Thesis document presented (or) as a partial requirement to qualify for the degree de: **Doctor in Biotechnology**

Supervisor (a):
Prof. Dr., Fabio Ancizar Aristizabal Gutierrez

Co Supervisor (a):
Prof. Dr., Diego Mauricio Riaño Pachón

Research Line:
Bioprocess and bioprospection - Natural Rubber

Universidad Nacional de Colombia
Sciences Faculty, Biotechnology PhD program
Bogotá, Colombia
2021

I want to thank God, my mother, my siblings Oscar, Laura, and Carolina, and my nephew Camilo. This project is irrefutable proof that everything you wish for can become true. Do not forget that dreams are achievable goals, which can only be accomplished through discipline and persistence. So, be strong and keep forward always

I also want to thank Daniel Tello for being my partner, my strength, and the motivation to accomplish this life project. All my love and gratitude to you because you came to my life to bring calm, peace, and happiness.

Acknowledgment

To Diego Mauricio Riano Pachon who has been my teacher and co-worker in this project. All my gratitude goes to him for solving in the most assertive way questions in the biological and bioinformatics fields, supporting my transition with patience and humanity. I also want to thank him for believing in me, giving me the liberty to build this project together. To the Laboratory of Computational, Evolutionary and Systems Biology and the center for Nuclear Energy in Agriculture from Sao Paulo University for the Computational source supplied to the development of my thesis project.

To Fabio Aristizabal for the unconditional support during my PhD program. For making the execution of this project easier during the setbacks and encouraging me to keep going until accomplishing my goals.

To Ibonne Garcia for the advice in the experimental part of this project and for the constant support to find the biological meaning of my results. For being my friend and my labmate in the last years.

To Daniel Croll and the Laboratory of Evolutionary Genetics in Neuchatel University Switzerland for their huge contributions in the structure and methodology of this project and for encouraging me to be a better researcher every single day of my PhD internship. To Lea, Sabina, and Leen for their contributions to the review of this manuscript.

To Ursula for her huge contribution to the analysis of this project. And for the unconditional support in the hard moments. Since I met you, you have been light in the difficult and dark paths, so, my achievements are also yours.

To Professor Jaime Robledo from Corporación para Investigaciones Biológicas - CIB, Unidad de Bacteriología y Micobacterias, as well as the researchers, Uriel Alonso Hurtado Paez and Nataly Álvarez Zuluaga, for the facility and the advice in nanopore sequencing technology.

To the Biotechnology Institute and the National University of Colombia for being my house for the last 17 years, I am deeply grateful to be part of this community. I want to thank to Diana Vinchira, Elizabeth Mendez, Jaqueline, Diana Pinzón, La parrita, Miriam, Roci, Mauricio Bernal, Yeimi Pizón.

To Colciencias and Colfuturo for the financial support and the fellowship administration.

Resumen

Título en español: Ensamblaje del genoma de alta calidad y genómica comparativa de *Pseudocercospora ulei* GCL012, el agente causal del tizón foliar sudamericano (SALB) en el árbol de caucho natural *Hevea brasiliensis*: hacia la predicción de componentes moleculares asociados con su patogenicidad y virulencia

Pseudocercospora ulei es el agente causal del mal suramericano de las hojas del caucho (SALB), la principal amenaza de *Hevea brasiliensis*, una especie nativa de la Amazonía que representa la fuente comercial de caucho natural. El cultivo de *H. brasiliensis* es una alternativa económica importante para los países latinoamericanos; sin embargo, las estrategias de control de la enfermedad SALB son ineficientes debido a la dificultad del fitomejoramiento tradicional en esta especie perenne y también a la falta de datos genéticos sobre este hongo. Nuestro principal objetivo fue comprender los mecanismos moleculares que gobiernan la biología básica y el proceso de patogenicidad de *P. ulei*. Para eso, secuenciamos, ensamblamos y anotamos su genoma a través de una secuenciación de escopeta de genoma completo usando lecturas largas y cortas. Construimos el genoma más grande de la familia Mycospharelaceae con 98,3 Mbps que comprende 214 andamios con un valor N50 de 2,8 Mbps y una integridad BUSCO del 97,5%. El genoma de *P. ulei* alberga 12'745 modelos de genes de los cuales 756 se clasificaron como proteínas secretadas y 113 fueron candidatos a efectores, incluidos 54 que presentan actividad potencial en el apoplasto del huésped. *P. ulei* exhibe un contenido notablemente reducido de CAZymes y de agrupaciones de genes asociados a metabolismo secundario con 216 y catorce asignaciones respectivamente. Este genoma tiene un contenido repetitivo excepcional, el 80%

del tamaño del genoma está ocupado por elementos repetitivos clasificados principalmente en Clase I revelando una expansión del tamaño del genoma a través de elementos repetitivos que podrían jugar un papel esencial en la adaptación fúngica ambiental y la evolución de los mecanismos de patogenicidad.

Palabras clave: Caucho natural, Mal suramericano de las hojas del caucho, Pacific BioScience, Nanopore, Elementos transponibles.

Abstract

English Title: High-quality genome assembly and comparative genomics of *Pseudocercospora ulei* GCL012, the causal agent of the South American leaf blight (SALB) in natural rubber tree *Hevea brasiliensis*: Towards the prediction of molecular components associated with its pathogenicity and virulence

Pseudocercospora ulei is the causal agent of South American Leaf Blight (SALB), the main threat of *Hevea brasiliensis* an Amazonia native species that represent the commercial source of natural rubber. *H. brasiliensis* crop is an important economical alternative to Latino american countries however, SALB disease control strategies are inefficient as a reason for the difficulty of traditional plant breeding in this perennial species and also for the lack of genetic data about this fungus. Our main goal was to understand the molecular mechanisms that govern the basic biology and pathogenicity process of this fungus. For that, we sequenced, assembled, and annotated its genome through a whole-genome shotgun sequencing using long and short reads. We built the biggest genome of the Mycospharelaceae family with 98.3 Mbps comprising 214 scaffolds with an N50 value of 2.8 Mbps and a BUSCOs completeness of 97.5%. The *P. ulei* genome harbour 12’745 gene models form which 756 were classed as secreted proteins and 113 were as effectors candidates including 54 presenting potential activity in the host apoplast. *P.ulei* exhibits a remarkably reduced content of CAZymes and secondary metabolism genes clusters with 216 and fourteen assignments respectively. This genome bears an exceptional repetitive content, 80% of the genome size is occupied by repetitive elements mainly classified in Class I revealing a genome size expansion via repetitive elements which could play an

essential role in the environment fungal adaptation and the pathogenicity mechanisms evolution.

Keywords: natural rubber, South American Leaf Blight, PacBio, Nanopore, transposable element.

Content

Content

Figure list

Pág.

Figure list

Table list

Pág.

Table list

1. Symbols and abbreviations list

Symbols with Latin letters

Symbol	Term
g	Grams
Kg	Kilogram
L	Liters
mg	Milligrams
mL	Milliliters
U	Unit
ng	Nanograms
Mbps	Mega base pairs
Gb	Giga base
Kd	Kilodaltons

Symbols with Greek letters

Symbol	Term
µL	Microliter
µm	Micrometer
µM	Micromolar
λ	Lambda

Abbreviations

Abbreviation	Term
NR	Natural Rubber
SALB	South American Leave Blight
CAZymes	Carbohydrate-Active enZYmes
TEs	transposable elements
QTL	Quantitative Trait Locus
RFLPs	Restriction fragment length polymorphism
AFLPs	Amplified fragment length polymorphism
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
DGEs	Differentially genes expressed
RACE	Rapid Amplification of cDNA Ends
M/PAMP	Microbe-Associated Molecular Patterns
PTI	Pathogen Triggered Immunity
PRR	Pattern Recognition Receptors
ROS	Reactive Oxygen Species
ETI	Effector-Activated Immunity
HR	Hypersensitive Response
CDWEs	Cell Wall Degrading Enzymes
GH	Glycoside Hydrolases
GT	Glycosyltransferases
PL	Polysaccharides Lyases
CE	Carbohydrate Esterases
CBM	Carbohydrate-Binding Modules
AA	Auxiliary activities
SMs	Secondary Metabolites
NRPS	Non-Ribosomal Peptide Synthetase
PS	Polyketide Synthase
NRP	Non-Ribosomal Peptides

PK	Polyketides
TCs	Terpene cyclases
MAPK	Mitogen-activated protein kinases
cAMP	Cyclic Adenosine Mono-Phosphate
ORF	Open Reading Frames
AP	Aspartic Protease
GAG	Capsid Protein
RT	Reverse Transcriptase
LTR	Long Terminal directed Repeats
Tase	Transposase enzyme
TIR	Terminal Inverted Repeats
HGT	Horizontal gene transfer
ZMW	waveguides a zero point
OLC	Overlap Layout Consensus
MHAP	MinHash Alignment Process
BOG	Best Overlap Graph
UTR	Untranslated Regions
EST	Expressed Sequence Tag
NRDB	Non redundant database
BUSCO	Benchmarking Universal Single-Copy Orthologous
CDS	Coding Sequence Regions
EVM	Evidence Modeler
SSP	Small Secreted Proteins

Introduction

Natural rubber is a commodity that presents insuperable physicochemical properties which position it as the primary raw material for the industrial production of tires, glovers and condoms (Yamashita & Takahashi 2020). More than 2,500 plant species produce natural rubber, from these, *Hevea brasiliensis* is the commercial source due to its high production rates (Lieberei, 2007). The genus *Hevea* is native to the Amazon river basin, but since 1876 it has been grown in east Asia, where it adapted with success. Currently, the leading natural rubber producers are Malaysia, Thailand, and China, supplying 92% of global natural rubber, while Latin American countries even though being the origin center are uncompetitive in rubber production and commercialization (International Rubber Study Group (IRSG, 2019). The main reason for the low natural rubber production in this region is the disease known as the South American leaf blight (SALB), which reduces the rubber production by up to 75% (Chee, 1976). The primary control strategy against SALB is genetic control by planting resistant clones. Breeding clones with durable resistance without pleiotropic effects on rubber yield remains challenging. Rubber plantations in East Asian countries are constituted mainly by SALB susceptible clones, hence the arrival of *P. ulei* to the Asian continent constitutes a very significant economic risk (Lieberei, 2007). Strengthening *Hevea* breeding programs by introducing pathogen-informed selection is essential.

The causal agent of SALB is the ascomycete fungus *Pseudocercospora ulei* (formerly known as *Microcyclus ulei*). Despite the agronomic importance of this pathogen, knowledge about the molecular bases governing its biological and

pathogenic behavior is minimal. The lack of molecular knowledge is associated with its presumed biotrophic lifestyle which makes difficult the propagules obtention by *in vitro* culture (Chee, 1978). *P. ulei* taxonomic classification also had been challenging, several changes had been performed from its first description (Ericsson & Hawksworth, 1993; Gasparotto et al., 1984; Schubert et al.2003), however, the first description based on molecular information were recently described classifying the different reproduction stages in a unique family (Mycospharelaceae) placing them in the *Pseudocercospora* genus and renaming as *Pseudocercospora ulei* (Hora et al., 2014). The above highlights the necessity of a broad span molecular study that contributes to solving the big questions about the biology of this important species.

This work presented in this thesis encompasses the nearly complete and annotated genome of *P. ulei* and a comparative study with other fungal pathogen species belonging to Mycosphaeracea family like the black sigatoka complex (*P. fijiensis, P. emusae, P, musae*). Here, We analyzed the genome structure in terms of genome architecture properties and pathogenicity candidate proteins. Based on RNA-seq assisted gene predictions, we predict effector genes, Carbohydrate-Active enZYmes (CAZymes) and secondary metabolite gene clusters to establish a resource to investigate molecular determinants of SALB.

Chapter 1 presents the state of the art regarding natural rubber crop, SALB disease, *P. ulei,* and the genomics as a tool to understand fungal pathogens. Chapter 2 describes the *P. ulei* genome structure discussing the main features of its architecture. This chapter elucidates the evolutionary relationships of *P.* in the Mycospharelaceae family and in the *Pseudocercospora* genus, which were based on orthologous gene components and four genomic loci, respectively. The function of the *P. ulei* genome is described in chapter 3. In this chapter we characterize the predicted gene functions, focusing on the pathogenicity-related proteins like CAZymes, effectors, secondary metabolites, and transposable elements (TEs). Finally, the collected information throughout genome assembly and functional

annotation will be adapted into a comprehensive model of the *P. ulei* and *H. brasiliensis* molecular interaction, given for the SALB disease development.

The present work has been submitted to genetic and molecular biology journal below the title “High-quality genome assembly of Pseudocercospora ulei the main threat to natural rubber trees” it was a work developed by the Laboratory of Computational, Evolutionary and Systems Biology, Center for Nuclear Energy in Agriculture - the University of SãoPaulo, 13400-970, Piracicaba, SP, Brazil, and the laboratory of Molecular Characterization, Biotechnology Institute - Universidad Nacional de Colombia, Bogotá D.C., Colombia. In collaboration with the laboratory of Evolutionary Genetics - Neuchâtel University, Neuchâtel, Switzerland.

Chapter 1

1. Conceptual Framework

1.1. Natural rubber crop as an economical alternative

Natural rubber is a secondary metabolite constituted by isoprene molecules assembled uniformly by 1-4 *cis* double bonds. Due to its organized chemical structure, this polymer presents unique physicochemical properties (Yamashita & Takahashi, 2020). Natural rubber's exceptional properties like high temperature and impact resistance, malleability, and resilience (Hayachi, 2009) consolidate it as the primary raw material in the tires industry, which consumes more than 75% of the total worldwide rubber produced (Yamashita & Takahashi, 2020). The remnant rubber is used to fabricate more than 40,000 medical devices as glovers and condoms (Mooibroek & Cornish, 2000).

Natural Rubber is naturally produced by more than 2500 plant species (Mooibroek and Cornish, 2000); however, the species *Hevea brasiliensis* Muell. Arg., also known as the Brazilian pará tree, is the industrial rubber source for high production performance (Venkatachalam et al., 2013). *Hevea* species belongs to the Euphorbiaceae family and is native to the Amazon river basin. Rubber tree was transferred to the East of Asia, where it had an excellent adaptation process, being extensively planted as an economical alternative (Lieberei, 2007).

The *Hevea brasiliensis* is a perennial species that presents an initial 5-year unproductive period but later reaches an economically useful life of about 50 years (Compagnon, 1985; Lieberei, 2007). Besides its essential use for society, the economic importance of this crop also lies in the commercial plantation supply chain and has an economic impact on more than 20 million people (Rivano et al., 2016).

1.2. South American Leaf Blight, the main constraint of Natural Rubber crop

The leading natural rubber producers are Malaysia, Thailand, Indonesia, and China, supplying 91.2% of global natural rubber, while Latin American countries have been uncompetitive in rubber production (IRSG, 2019). The main reason for the low rubber production in Latin American countries is the South American leaf blight (SALB) disease caused by the ascomycete fungus *Pseudocercospora ulei* P. Henn. Von Arx. The SALB disease affects the trees' photosynthetic capacity, reducing rubber production by up to 75% (Chee, 1976). The SALB disease was first described at the beginning of the 20th century, and its most severe outbreak was reported in 1986 in Brazil and caused losses of about 150 thousand hectares (Lieberei, 2007). Currently, the SALB has been registered in more than fifteen countries all within South America (Edathil, 1986) (Figure 1-1). Despite the phytosanitary control attempts, this disease remains the primary threat from natural rubber crops in this region (Gasparotto et al. 1997).

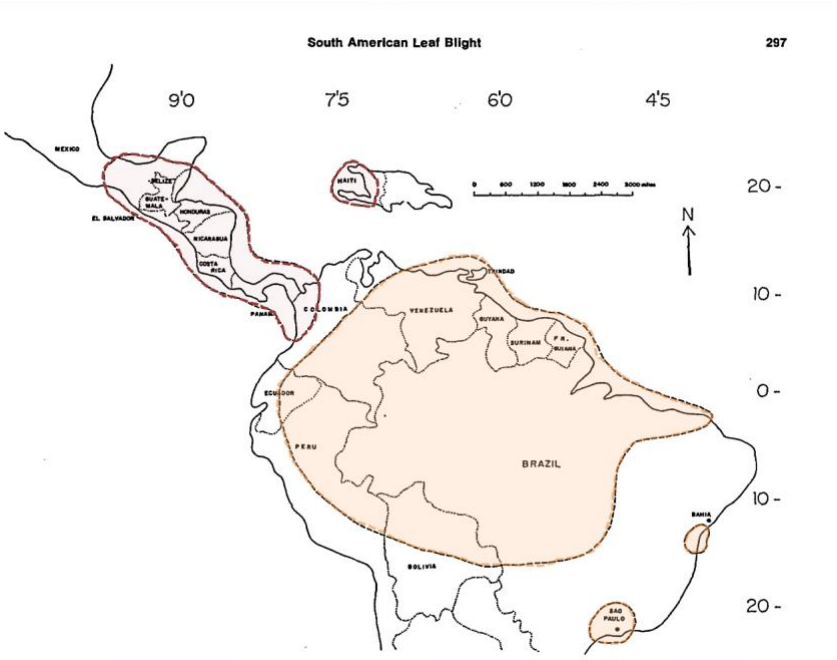


Figure 1-1. South American Leaf Blight (SALB) disease distribution, figure adapted from Edathil, 1986

Regarding signs of the SALB disease, the infection target tissue is immature leaflets, which once had been colonized present small sporulation spots recognized by its greenish-brown fluffy appearance (García et al., 2006). Once the leaf maturity is achieved, the described lessons take a bulky black appearance giving rise to the stroma mainly visible at the leaflet's upper part (Chee and Holliday, 1986, cited by Serie Técnica, CONIF, 1997). Finally, the density of the stromatic lesion prevents the photosynthesis process natural course, and the leaves fall after reaching full maturity (Gasparotto et al.,1984). Successive rounds of defoliation could result in tree exhaustion and its subsequent death (Lieberei, 2007). Young plantations have a higher SALB incidence due to infection of new sprouts and immature leaves that result from pruning. Likewise, productive plantations are at permanent risk because of the natural defoliation process, where young leaves replacing the fallen leaves become vulnerable to pathogen invasion.

1.3. Combating SALB disease, efforts and challenges

Several approaches have been adopted for SALB phytosanitary control in both clonal gardens and plantations. Chemical alternatives using fungicides such as mancozeb and benomyl are efficient to control leaf infection and pathogen sporulation in clonal gardens (Rocha et al.,1978; Gasparotto et al., 1984). Unfortunately, the application of fungicides in adult plantations is challenging because the trees can reach a height of up to 20 meters, and hence application of fungicides significantly increases management costs (Nilton, et ál., 1992). Strategies based on biological control have similar caveats.

An alternative control strategy has been identifying “escape zones” that present suitable climatic conditions for the tree development but not ideal for the pathogen infection cycle (Camargo et al,.2003). Although the latter strategy has been widely used in Latin American countries, it is only a temporary solution because the pathogen can adapt over time to survive in previously identified escape zones.

Moreover, the climatic change phenomena are rapidly affecting the already defined escape zones, making them suitable for developing the pathogen.

Similar to other crops, genetic control has been one of the most promising alternatives for SALB disease mitigation (Gonçalves et al., 1997). *Hevea* cultivars with resistance potential against SALB disease have been generated through *Hevea* Breeding Programs (Junqueira, et ál., 1992). Breeding assays date back to 1927 (Rivano et al. 2013), but it was not until the beginning of the 21st century that the first genetic markers associated with SALB resistance were identified (Lespinasse et al. 2000a; Lespinasse et al. 2000b). Lespinnasse and collaborators generated the first genetic linkage map for *H. brasiliensis* from the F1 progeny between the resistance accession RO-38 a *H. brasiliensis* × *H. benthamiana* interspecific hybrid clone and the susceptible accession PB-260 (Lespinasse et al. 2000a). The saturation of this map with 301 Restriction fragment length polymorphism (RFLPs), 388 Amplified fragment length polymorphism (AFLPs), 18 microsatellites, and 10 isozymes carried on to the identification of 18 linkage groups revealing the diploid nature *H. brasiliensis* tree (2n = 36 chromosomes). The described results were the building block for the systematic identification of Quantitative Trait Locus (QTL) markers associated with SALB disease resistance (Le Guen et al. 2006).

The molecular screening of genetic linkage parentals has allowed identifying the M13-lbn (g13 linkage groups) Quantitative Trait Locus (QTL), which explains until 90% of the phenotype variability in controlled conditions (Le Guen et al., 2003). Further, the identification of five QTL (g11 and g12 linkage groups) suggests that SALB's genetic resistance is more oligogenic than polygenic (Lespinasse et al., 2000a; Le Guen et al., 2003; Le Guen et al., 2008). Latino American *Hevea* breeding programs by their part, continuously working to generate good performance clones with SALB resistance traits. To date, these programs are focused on assessing the genetic potential of derived clones from the Brazilian, Ecuadorian and Colombian programs at several time scales (Rivano et al., 2010;

Sterling et al., 2018). However, considering the variability of climatic conditions in suitable areas for planting rubber crops of Latinoamerican countries, the necessity to evaluate the rubber production yield and resistance traits in environmentally differential zones is a matter of fact.

Compared with other species breeding programs, *Hevea* breeding programs advances slowly due to natural rubber tree perennial nature (Priyadarshan et al., 2009). The evaluation period of *Hevea* breeding programs could reach 25 years (Costa et al., 2000). Although the scientific community keeps working to reduce these evaluation times (Kalpani, Withange & Palihakkara, 2020), it remains the most significant constraint of the breeding practices in this crop. Besides, the narrow genetic base of the *Hevea* spp. cultivated species represent another significant drawback of *Hevea* breeding programs, this is because most of these clones are proceeding from a few seeds rescued in Wickham collection (Oghenekome, 2004). The regulations in the transfer of genetic materials make it challenging to introduce alleles from wild relatives. Notwithstanding the described constraints, genetic control by planting resistant clones is still the more promising strategy for SALB control. The described advances are highly relevant to the worldwide rubber future, mainly to East Asia countries whose plantations are constituted by SALB susceptible clones. Thereby, the transfer of *Hevea* clones with SALB resistance traits could be plausible, avoiding the devastating effect on the worldwide economy that would cause the arrival of *P. ulei* to the Asia continent.

The assessment of non-traditional plant-breeding methods is elemental for effective control of SALB disease. Alternative solutions based on genetic engineering are key to confer durable resistance without pleiotropic effects to commercial cultivars in a shorter period. However, the selection of potential mutational targets to confer resistance against SALB through modern technologies such as CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) requires prior in-depth knowledge of the main molecular actors of the *H. brasiliensis* - *P. ulei* interaction. Besides, strengthening *Hevea* breeding programs by introducing pathogen-

informed selection remains essential. Hence, this work aims to generate *P. ulei* genomic resources performing a detailed understanding of the molecular bases of the basic biology and the pathogenicity determinants that govern SALB disease development. The following sections describe the current knowledge of genetics, biology, and pathogenicity of *P. ulei*.

1.4. The current knowledge about the molecular determinants that govern the SALB disease

The major knowledge contributions about molecular determinants that regulate SALB disease are from the host's perspective. The first *H. brasiliensis* draft genome was released in 2013 (Rahman et al.,2013), and a complete version was further finished in 2016 (Tang et al., 2016). The genome sequencing availability enable the first systematic characterization of resistance genes in rubber trees; however, the knowledge from the pathogen perspective is still a study subject. Similarly, two transcriptomic studies of the *H. brasiliensis* and *P. ulei* interaction have been published however they are focused to characterize the host response during SALB disease (Hurtado et al., 2015; Angelo et al., 2020). Although SALB interaction transcriptomes could help to learn about both host and pathogen, the lack of authentic information about the fungus hinders studying the disease from the pathogen perspective.

From the plant perspective, Hurtado and collaborators reported the first transcriptome of the interaction of *P. ulei* with a partial resistant *H. brasiliensis* clone identifying differentially genes expressed (DGEs) in the early SALB development stages (0 and 48 hpi). This study identified 86 differentially expressed transcripts; 90% of them explain plant defense response mechanisms, including cell wall remodeling, chitin degradation, and salicylic acid pathway activation (Hurtado et al., 2015). Likewise, the second differential gene expression analysis of SALB interaction, revealed plant immunity expression shifts pointing out that processes like ubiquitination, cytoplasm detoxification, cell wall reinforcement, calcium, hyper sensible response, and cyanogenesis could be active during SALB disease

development (Angelo et al., 2020).

From the pathogen perspective, some studies provide limited information about the invasion and spread processes. Fluorescence observations of *P. ulei in situ* colonization revealed the penetration process is mediated by a specialized structure called appressorium (Hashim, Chee, and Duncan, 1978). However, appressorium seems not to be mandatory for the invasion process, and its formation dependent on interaction in the pre-penetration phase (Lieberei, 2007). It is also known that *P. ulei* is tolerant to cyanogenic glycosides secreted by the host during the early stages of infection (Lieberei et ál., 1983). This is a metabolic adaptation to the high production of cyanogenic compounds by rubber plants and the degradation of cyanide compounds by *B*-glucosidase activity (Lieberei et ál., 2006).

The lack of a reference genome for *P. ulei* has been a significant constraint to determine the pathogenicity mechanisms involved in SALB. The natural rubber research group of Universidad Nacional de Colombia recently carried out several studies, contributing to the characterization of *H. brasiliensis* - *P. ulei* interaction (Hurtado et al., 2015). For example, this group partially identified and sequenced an effector protein termed TDFMu1 through the analysis of cDNA-AFLPs between resistance and susceptible clones inoculated with virulent *P. ulei* strain. Sequencing of differentially expressed bands enable to obtain a partial sequence of TDFMu1 highly similar to ECP-5, which is an apoplastic effector of the *Cladosporium fulvum* fungus (Butler & Jones 1949). Further studies attempted to obtain the complete sequence of this *P. ulei* effector candidate by a RACE (Rapid Amplification of cDNA Ends) approach, but it was impossible to extend the 3' extreme, and this protein is still unresolved.

1.5. *Pseudocercospora ulei* a poorly studied pathogen

The ascomycete fungus *Pseudocercospora ulei* (former *Microcyclus ulei*) is the main threat to natural rubber commercial crops. Despite the agronomic importance

of this pathogen, molecular knowledge about the basic biology and pathogenicity is limited. This lack of knowledge is mainly associated with the *P. ulei* presumptive biotrophic lifestyle that hinders its *in-vitro* manipulation requiring long incubation periods (more than 40 days) and high nutritive culture media (Chee, 1978, Lieberei, 2007). Although the *in vitro* culture difficulties, the *P. ulei* morphological characterization is well defined, and several reproduction forms have been identified by microscopy observation. *P. ulei* exhibits a teleomorphic phase termed *Microcyclus ulei* (P. Henn) v. Arx, sin. *Dothidella ulei* P. Henn (Gasparotto et al., 1984) and two anamorphic forms (deuteromycota): the pycnidial phase *Aposphaeria ulei* (P. Henn) and the conidial phase *Fusicladium Microsporum* Kuiper (Schubert et al. 2003) which is the unique phase possible to propagate in the laboratory and the responsible structure of SALB dissemination.

In reference to the *P. ulei* taxonomic classification, it has been challenging, having multiple changes from its first description. The first phylogenetic classification of *P. ulei* was performed based on morphological traits, placing each reproduction form in a taxonomic group as a different species (Hennings, 1904). Ten years later, the three reproductive forms were linked in one single fungus termed *Melanopsammopsis heveae* (Cayla & Petch, 1914). It was only until 1962 when stroma morphology observations led the *M. heveae* classification to the *Microcyclus* genus terming it *Microcyclus ulei*. Thirty years later, the *Microscyclus* genus was included in the Mycospharelaceae family (Ericsson & Hawksworth, 1993); however, a second stroma morphology study transferred *M. ulei* to the Planistromellaceae family and Dothideales Order (Lumbsch & Huhndorf, 2010). Finally, the three *M. ulei* forms were included in a phylogenetic study based on the partial sequences of LSU_ rRNA, actin and, elongation factor genes, whose results allowed putting it back to the Mycospharelaceae family changing them to *Pseudocercospora* genus terming them as *Pseudocercospora ulei* (Hora et al., 2014).

The *P. ulei* nutritional metabolism has been widely debated. Some authors have described it as an obligate biotroph (Lieberiei 2006; Lieberiei, 2007), others describe

that *P. ulei* presents a hemibiotrophic nutritional lifestyle (Sterling & Melgarejo, 2018). Even, the recent publication of Furtado and collaborators describes this fungus as a necrotrophic pathogen (Furtado et al., 2020). Although the evidence of a necrotrophic phase in SALB disease is still unclear, there is no doubt that *P. ulei* requires the host alive to complete its life cycle. In consequence, *in vitro* manipulation of *P. ulei* is extremely challenging and the propagules obtention could take until 50 days.

As the disease cycle is usually studied through *in situ* assays or using specialized stains and microscopic observations, the *P. ulei* lifecycle is very well characterized, as is described in the following lines. *P. ulei* life cycle starts with the release of ascospores, which represent its teleomorphic phase and acts as the primary inoculum of the disease. Ascospores are released from mature leaves that have fallen from previous infections as they are contained in a mycelium agglomeration termed stromata that are the causal reason for the *Hevea* trees defoliation (Gasparotto et al., 1984). The ascospores are released under specific temperature and relative humidity conditions and subsequently transported by rain and winds to the abaxial part of other young leaves. The adhesion to the target tissue and the subsequent germination of ascospores occurs within the next 12-24 hours, followed by the penetration process in which the fungus could develop a penetration structure known as appressoria (Chee et al.,1986). Thereby, the fungus extends the primary hypha inside of host tissue invading the leaves and causing the disease symptoms in a time window of 6-12 days post-infection. Once the fungus is disseminated inside the host it is possible to identify the fungal signs like sporulated lesions. The fungus sporulation leading to the conidia release, conidia are carried on by wind or rain being able to infect not only the closest healthy leaves but also those belonging to far plantations. Finally, clusters of disseminated conidia from stromata in newly infected leaves, sheltering new ascospores that will be released giving rise to a new disease cycle (Chee et al. 1986).

The pathogenicity knowledge of *P. ulei* has been limited to assessing the disease symptoms' aggressiveness on a scale of rubber tree accessions with differential resistance performances. Thus, has been evidenced the existence of races or strains with differential virulence profiles. For instance, the *in situ* evaluation of the sporulation intensity and the stromata density, enable to distinguish more than 36 pathogenic races, suggesting that *P. ulei* presents high intraspecies genetic variability (Chee and Holliday 1986; Mattos et al., 2003). Unfortunately, only one study of genetic variability is currently available, in which eleven microsatellite markers showed polymorphism across a limited population of six strains (Le Guen et al., 2004). As detailed in the next section, the molecular basis of the *P. ulei* pathogenicity remains mostly unknown.

1.6. Looking for new control disease strategies. A matter of two perspectives, host and settler

Fungi are ubiquitous microorganisms that establish symbiotic relationships with other living organisms to accomplish their heterotrophic nutrition. Although this microbial group is highly diverse, only a few of them (10%) generate damage in its host causing disease (Blackwell, 2011). Even though being a low percentage, pathogenic fungi are the main threat to commercial crops, causing significant economic losses, and representing a prominent risk in human feed security (Horbachab et al., 2011).

In order to complete their pathogenicity cycle, fungi have evolved sophisticated molecular mechanisms to manipulate the host circumventing the defense response activation (De Jonge et al., 2010; Hernandez et al., 2017). Consequently, the plant exhibits elaborate defense mechanisms to confine the pathogen to the invasion site, avoiding its spread and subsequent disease development. The host-pathogen competition mediated by a bipartite evolution to ensure survival and perpetuate fitness is called "the arms race" (Stahl and Bishop, 2000; Boller and Yang, 2009).

Understanding this dynamic interaction from both perspectives (pathogen and

plant) is essential to the new disease control strategies development. In this sense, the identification of SALB resistance genes would be highly valuable since they could be new *Hevea* breeding targets. Likewise, the *P. ulei* pathogenicity determinants knowledge contributes to strengthening *Hevea* breeding programs by introducing pathogen-informed selection, and besides in the new fungicide targets discovery.

Part of our motivation to perform this project is to balance SALB disease knowledge from the pathogen perspective. However, considering the previously described, we will briefly describe the plant immunity bases, and then, we will focus on describing the main fungal pathogenicity mechanisms studied in model fungi.

1.7. The host side, turning on the alarms to prevent the health breakdown.

The plant's defense mechanisms are divided into constitutive and inducible. Constitutive mechanisms are aimed to prevent the insects and microbial pathogens invasion and include preformed physical barriers, represented by cell walls, cuticle waxes, and plant bark. Inducible mechanisms, in turn, include the production of harmful substances for pathogenic microorganisms such as toxins and microbial wall degradation enzymes (Freeman & Beattie, 2008).

The inducible immune response in plants has two steps. The first is a non-specific early recognition response based on Pathogen or Microbe-Associated Molecular Patterns (PAMP or MAMP), and it is known as Pathogen Triggered Immunity (PTI) (Figure 1-2). The goal of this response is to prevent the pathogen entry and its subsequent proliferation (Zipfel and Felix 2005). To recognize the intruder microbes, plants have membrane receptors termed Pattern Recognition Receptors (PRRs) to perceive PAMPs or MAMPs (Zipfel 2008) (Figure 1-2 a). PAMPs are highly conserved molecules, essential for pathogen survival, and usually located in the microorganism's outer layer. The most studied PAMPs are flagellin and lipopolysaccharide for bacteria and chitin or fungi cell wall glucans (Chinchilla et

al.,2006; Pusztahelyi, 2018). This unspecific recognition triggers the gene expression related to the vegetal wall strengthening activating the callose and lignin deposition or could result in the release of antimicrobial compounds as Reactive Oxygen Species (ROS) and wall pathogen degradation enzymes (Jones & Dangl 2006).

A compatible host-pathogen interaction in PTI response results in successful colonization, and the pathogenic agents can take a step forward to proliferate inside the host. For this purpose, pathogens inject proteins known as effectors or Avr proteins (Figure 1-2 b). The effectors are virulence factors whose primary function is to counteract the plant defense response promoting the disease development (Cross, 2008; Lo Presti et al, 2015). Those proteins trigger the second innate immunity line termed the Effector-Activated Immunity (ETI). During this recognition, a resistance Protein (R) recognizes an effector protein (Jones and Dangl, 2006; Chisholm et al., 2006). This immediate recognition defines a gene-by-gene interaction (Flor, 1971) (Figure 1-2 b). This specific plant response leads to cellular death, known as Hypersensitive Response (HR), which is an aggressive mechanism to lockdown the pathogen growth (Figure1-2 d) (Zipfel and Felix 2005; Jones & Dangl 2006).

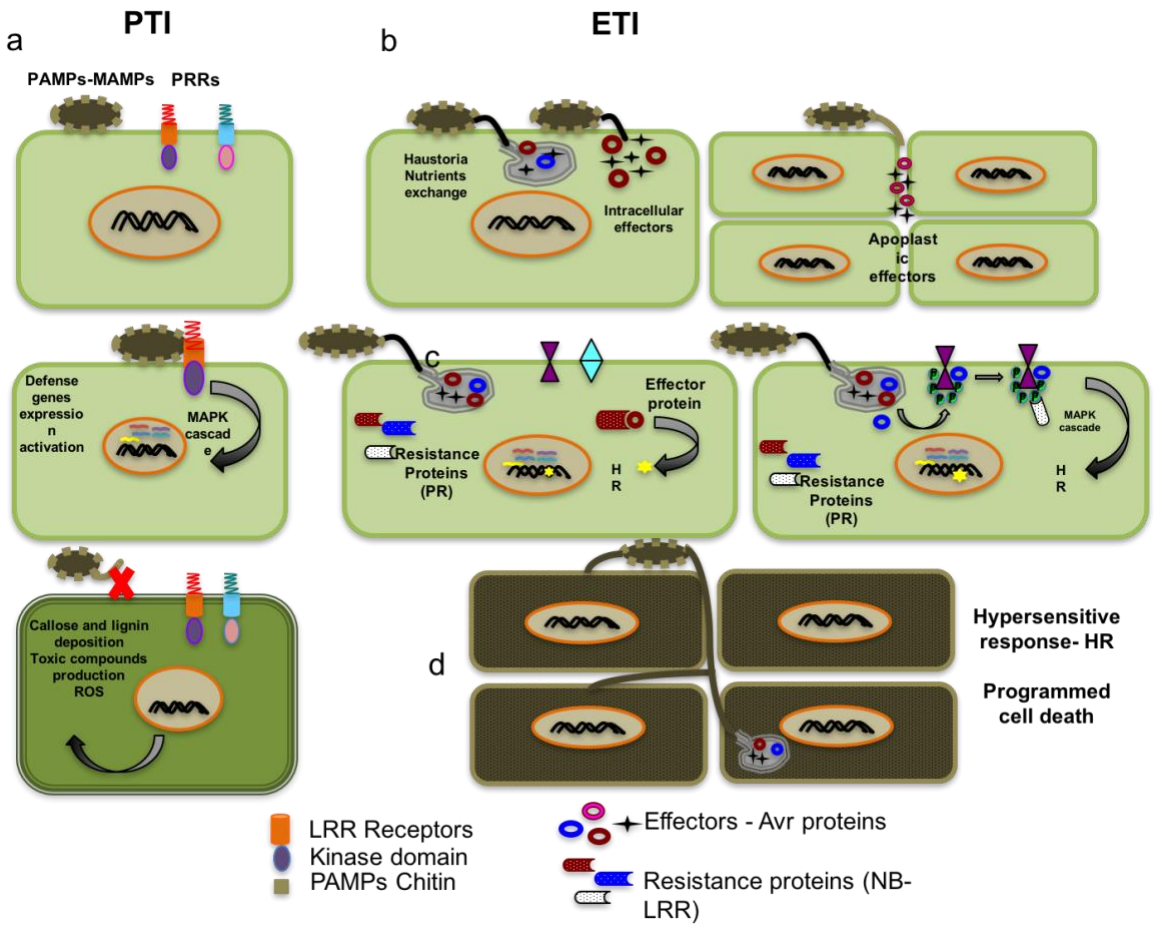


Figure 1-2. Plant immune system, source of this work

1.8. Fungal phytopathogens, a group of stealthy invasors

The first obstacle that fungal pathogens must deal with in an infectious process are the host physical barriers (first line of defense) represented by the cell wall and the cuticle. Vegetal cell-wall is a rigid structure is composed in more than 90% by polysaccharides (Houston et al., 2016), while, the cuticle is composed of fatty acids (cutin) linked by ester-type bonds (Kolattukudy, 1985; Holloway, 1982; Engelsdorf et al., 2017). The cuticle confers hydrophobic properties to the vegetal tissues' surfaces in order to protect them from insects and other pathogens. Therefore, pathogens invasion strategies are focused on establishing a strong contact with the host before starting the cell wall degradation process.

The adhesion is then the first step in a fungal infection and takes place after recognizing the host target tissue (Geoghegan, 2017). This process is carried out in two phases: In the first phase, the fungal primary inoculum establishes direct contact with the host cuticle, which is mediated by the secretion of adhesion proteins known as hydrophobins (Wang et al., 2010). In the second phase, the initial attachment is fortified by producing a natural glue (extracellular matrix) during the spore dormancy stage (Figure 1-3 a). This process strengthens the spore adhesion to the host cuticle, avoiding its shedding by wind or humidity. (Epstein & Nicholson, 2016).

The hydrophobins are small cysteine-rich peptides widely conserved in fungi (Blango, Kniemeyer & Brakhage, 2019). These proteins are located in the outer layer of the primary hypha and present amphipathic properties that favor the fungal invasion process (Linder et al., 2010). Besides their role in adherence, hydrophobins play a crucial function triggering the formation of penetration specialized structures (Talbot et al. 1993) once the host hydrophobic surface is detected by the fungus (Beckerman & Ebbole 1996). Additionally, the hydrophobin's role in the control of the host's immune response activation is widely described. The hydrophobin's structural organization in amphipathic monolayers appears to be key to avoid the plant defense elicitor's release during the penetration process (Bayry et al., 2012).

Once the spores are attached to the host, the pathogen needs to penetrate the cuticle and subsequently the plant cell wall to achieve a successful infection. To this, phytopathogenic fungi secrete Cell Wall Degrading Enzymes (CWDEs) aimed to soften the host cell wall (Figure 1-3 c). Beyond degrading cutin polymers, these enzymes can degrade other cell wall compounds like pectin, cellulose, lignin, and polysaccharides (Zhao et al., 2013). Accordingly, early stages of fungal diseases are characterized by a broad secretory activity characterized by production of CWDEs with peptidases, lipases and CAZyme activities.

The CAZymes are responsible for the degradation, synthesis, and modification of complex carbohydrates and glycoconjugates (Cantarel et al, 2009). To date, CAZymes are classified based on their catalytic modules in glycoside hydrolases (GHs), glycosyltransferases (GTs), polysaccharides lyases (PLs), and carbohydrate esterases (CEs) being the most abundant GHs and CEs categories (Zhang dbcan 2018). Additionally, there are two additional categories corresponding to carbohydrate-binding modules (CBM) and auxiliary activities (AA) (Zhang et al.,2018). Plant-pathogen fungi exhibit a massive CAZymes diversity in number and composition, this diversity correlates with the fungal nutritional lifestyle (Contier et al.,2012; Zhao et al, 2013, Lyu et al.,2015). Thus, fungal pathogens with necrotrophic metabolism exhibit an invasion strategy characterized by a plenty secretion of CAZYmes (Rodriguez et al., 2018). Contrary, biotrophic fungi present a limited hydrolytic activity, represented by a typical low content of coding genes of CWDEs (Kubicek et al, 2014; Albera, 2016). An attenuated degradation of the vegetal wall components in biotrophs fungi prevents the release of vegetal immune response elicitors like cutin monomers which acts as Damage-associated molecular patterns (DAMPs), activating of the first vegetal defense line (Bigeard et al., 2015 & Lo Presti et al., 2015).

In order to complement the colonization process, necrotrophic and hemibiotrophic fungi strengthen the host tissue damage secreting Secondary Metabolites (SMs). SMs are natural compounds produced by many organisms to counteract certain conditions or critical environments (Scharf et al., 2014). Fungi secrete a broad spectrum of SMs, including mycotoxins, pigments, siderophores, and antibiotics (Wang et al., 2014; Khaldi et al., 2010). These compounds have specific roles in pathogenicity (Scharf et al., 2014) as well as in the interactions with other species, acting as mediators of the nutrition process, and interspecies communication (Macheleidt et al., 2016).

Fungal SMs are produced by Non-Ribosomal Peptide Synthetase (NRPS) and Polyketide Synthase (PKS) enzymatic modules or independently by enzymes type prenyltransferases and dimethylallyl tryptophan synthases (Macheleidt et al., 2016). The SMs structure is determined by the number, organization and type of enzymatic modules involved in their synthesis. This last defines the SM classification into Non-Ribosomal Peptides (NRP), Polyketides (PK), Terpene cyclases (TCs) and hybrid complexes (Collemare et al., 2014). The SMs synthesis modules' combination possibilities give place to the production of compounds with wide chemical diversity emphasizing the biotechnology potential of these substances (Keller et al.,2005).

Fungal invasion process does not always compromise the host cell wall damage, alternatively, fungi can reach the host tissue through the plant stomata, minimizing the energy investment. In this approach, fungi can redirect their primary hypha detecting the opening stomata by a mechanism known as thigmotropism (Tudzynski and Sharon, 2003).

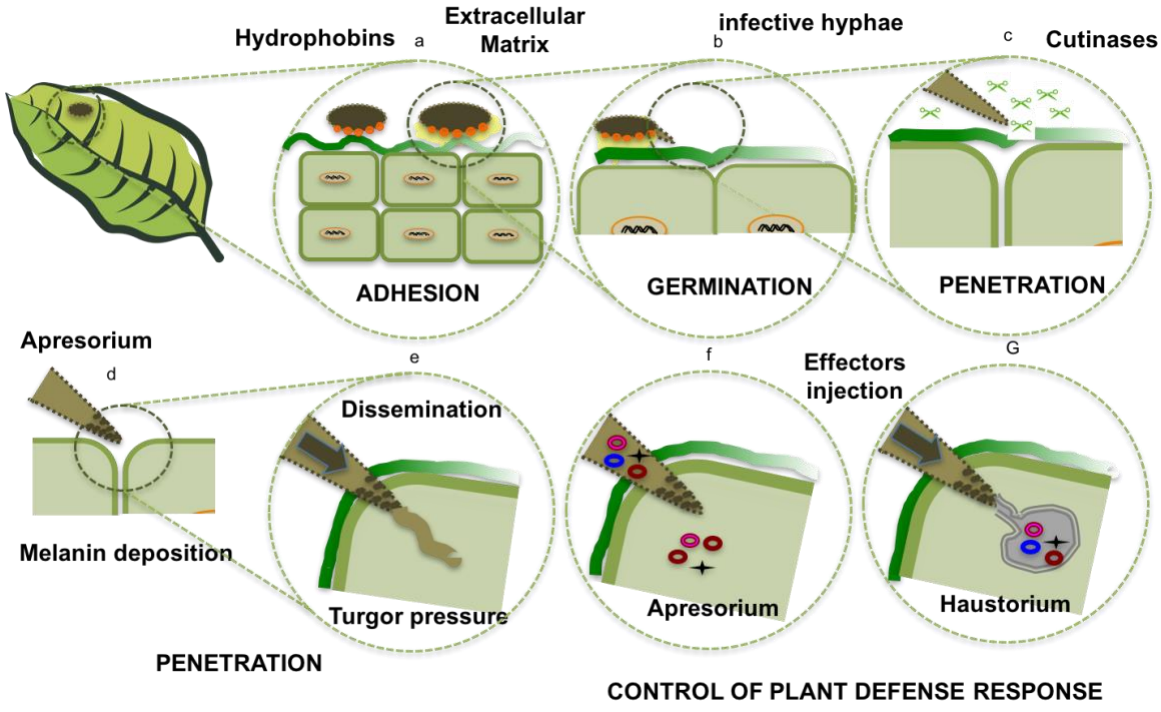


Figure 1-3. Pathogenicity mechanisms of biotrophic fungi, source this work.

1.9. Penetration phase, a fungal bet to win in the disease game

In a successful infection process, fungal pathogens establish an intimate contact with the host cells with two proposes: i) to accomplish an optimal nutrition process and ii) to trigger molecular manipulation mechanisms preventing the natural course of the plant immune response. The penetration phase is preponderant to the disease development and is defined by the formation of specialized penetration structures during an intense molecular activity that mediates the entry and spread fungal strategies.

Penetration structure formation is activated by the host surface hydrophobicity and plant cutin monomers released as a result of the colonization process (Bigeard et al., 2015). The perceptions of these signals by the pathogen leads to the activation of enzymes participating in the signaling cascades of MAPK (Mitogen-activated protein kinases) and cAMP (Cyclic Adenosine Mono-Phosphate) that are elemental in cell reproduction and differentiation that gives rise to appressorium formation (Ryder & Talbot, 2015). The appressorium is a penetration structure formed by the infective hypha distal cell specialization (Ryder & Talbot, 2015). The maturation of this structure involves remodeling of the distal cell cytoskeleton by rearranging actin molecules (F-actin) to generate a "spearhead" that will facilitate cuticle puncture (Talbot, 2003). The differentiation process of the infective hypha to appressorium occurs by morphological adjustments. First, the hypha tip is strengthened through the deposition of melanin, a high molecular weight pigment composed of phenols which confers resistance properties to the appressoria (Jacobson, 2000) (Figure 1-3 d), due the melanin role in the appressoria development it is considered a fungal pathogenicity factor (Langfelder et al., 2003; Scharf et al., 2014). Second, the fortified hypha is extended to the plant tissue's intercellular space by turgor pressure application (Howard et al.1991) (Figure 1-3 e).

Fungi appressoria formers can also assist the penetration process with CWDEs, but

their activity is lower than non-appressoria forming fungi (Casadevall, 2007). Besides enabling a more subtle penetration process, specialized penetration structures such as the appressorium and the haustorium facilitate the delivery of effector proteins into the host (Figure 1-3 f-g) (Kleemann et al., 2012; Doehlemann and Hemetsberger, 2013; Lo Presti & Kahmann 2017). The haustorium formation has been widely described in obligate biotrophic fungi such as *Uromyces viciae-fabae* and *Blumeria graminis f.sp hordei* (Lo Presti et al., 2015). The haustorium is a nutritional structure developed in the host's mesophilic cells (Garnica et al., 2014). Its formation has been associated with the penetration process of biotrophic fungi; however, it does not imply the host cell wall's penetration. This fungal structure is constituted by establishing a separation zone formed by the fusion of the membranes of the fungus with those of the host (Szabo & Bushnell, 2001). This area represents a suitable space for the nutrition of the pathogen, mainly for the assimilation of bio-elements such as amino acids and carbohydrates and facilitating the suppression of the defense response mediated by the secretion of effectors proteins (Staples, 2001) (Figure 1-3 g).

If the colonization process is successful, biotrophic phytopathogen intensifies the mechanisms to avoid plant immune system activation. For that, these invaders secrete effector proteins during the biotrophic nutritional period. The effectors are fungal secreted proteins that counteract the plant defense response promoting the disease development (Cross, 2008). Fungal effectors exhibit a conserved structural organization that comprises a small size (200-400 amino acids), cysteine residues abundance in their sequence, and signal peptide presence at its N-terminus, followed by multiple conserved domains at its C-terminus (Figure 1-4) (Stergiopoulos and de Wit, 2009; Sonah, Deshmukh, & Bélanger, 2016). Although effector proteins are highly variable, the abundance of cysteine residues and the signal peptide presence are indispensable for folding, stability, and secretion, respectively (Doehlemann and Hemetsberger, 2013).

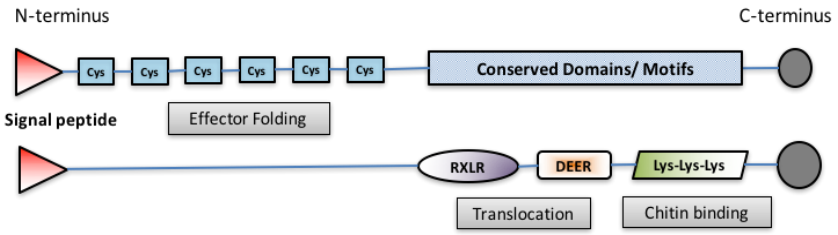


Figure 1.4. Fungal effector structure

Effectors are classified according to the place where they are excreted as apoplastic and cytoplasmic effectors (Figure 1-2 b) (Stergiopoulos and de Wit 2009; Shirasu, 2015). However, the translocation mechanisms to the host's interior are still unknown (Kamoun, 2014). The presence of conserved domains and protein motifs at the C-terminus end is one of the first advances in the knowledge of effector translocation mechanisms.

Concerning the effector's function, it is associated with its location. Thus, apoplastic effectors present proteolytic activity to inactivate CWDEs secreted by the plant (O'Connell, & Panstruga, 2006). Apoplastic effectors can also kidnap chitin monomers released during the invasion, this activity is mediated by chitin recognition for LysM motives present in this group of fungal effectors (Kombrink et al., 2017). Cytoplasmic effectors activity by its part are aimed to modulate the host biology at the metabolic level (Djamei et al 2011), either by redirecting central pathways to those less effective pathways with the object of decreasing toxic substances produced by the plant (Tanaka, Han & Kahmann, 2016), block compounds key in MAPK cascades activation, plant ubiquitination systems, hormonal signaling, and gene transcription systems (Asai & Shirasu, 2015).

1.10. Genomic as a modern tool to understand fungal pathogens

The development of high-throughput sequencing technologies represented a breakthrough in genome assembly projects as they enabled a dramatic reduction in

DNA and RNA sequencing costs (Goodwin et al., 2016). Nowadays, the assembly of a genome reference sequence for a species is a building block for most primary and applied research. Fungi are the major sequenced eukaryotic group (Jun-Ma & Fedorova 2010; Galagan et al., 2005). The extensive surveys of fungal genomes have revealed the diversity comprehended in their genome architecture, describing variable rates for genome size, gene content, transposable element (TEs) occupancy (Stukenbrock & Croll, 2014; Möller and Stukenbrock, 2017; Mohanta & Bae, 2015).

The genome architecture variability previously described has been remarkable in pathogen fungi, which present exceptional genomic plasticity (Möller and Stukenbrock, 2017). Fungal pathogens' genomes evolve rapidly to adapt to the host's harsh environment, giving rise to genetic changes that often confer fitness-related traits. TEs are the control agents of the fungal genome organization and gene content modulation (Daboussi & Capy, 2003). These versatile elements are defined as genetic units able to move to new host genome locations such as exons, introns, or regulatory regions (Kempken & Kück, 1998). The transpositions events could have neutral, positive, or negative effects on the host genome, driving to loss or gain genes, duplications, copy number variation, and horizontal gene transfer alterations (Stukenbrock & Croll, 2014).

TEs are classified based on their dispersion mechanisms in two big categories. The class I elements or retrotransposons, which transpose by RNA intermediate coding its own reverse transcriptase, and class II or transposons that transpose at DNA level following an excision and insertion cycle. (Muszewska, Hoffman & Grynberg, 1998; Rebollo, Romanish & Mager, 2012). Besides these two classes can be hierarchically subdivided into different subclasses, orders and superfamilies depending on their structural features (Wicker at al., 2007). The retrotransposons structurally compromise open reading frames (ORF) for viral structural proteins transcription needed for their transposition i.e., aspartic protease (AP), capsid protein (GAG), reverse transcriptase (RT), RNAse H (RH), a DDE integrase (INT).

These ORF could be or not flanked by Long Terminal directed Repeats (LTRs), originating the LTRs and non-LTR orders. Alternatively, the transposons exhibits a different structure that comprehend a transposase enzyme (Tase) which mediates their transposition flanked by Terminal Inverted Repeats (TIR) (Daboussi & Capy, 2003; Wicker at al., 2007).

The mechanisms that govern the adaptive evolution of fungal genomes have been studied through comparative genomics in fungal populations. These studies have identifying that the TEs have a transversal role in the generation of genomic changes either large and small scale (Priest et al., 2020). Chromosomal rearrangements, changes in the number of the set of chromosomes or part of them (aneuploidy and polyploidy, accessory chromosomes), and allelic transmission by Horizontal gene transfer (HGT) constitute large-scale genomic changes (Stukenbrock & Croll, 2014; Priest et al., 2020). In contrast, small-scale genomic changes are represented by single-base polymorphism, deletions, or insertions generated by accelerated hypermutation rates or defects in the DNA reparation mechanism. These latest could be directed to specific genomic regions, e.g., intron-exon junctions, intergenic sequences, and zones rich in pathogenicity genes (Priest et al., 2020).

Fungal genomics advent also has revealed the existence of fungal genome compartments consisting in high density gene regions that are often flanked by TEs (Möller and Stukenbrock, 2017). Due to the TEs activity, these regions exhibit higher evolve rates than other genomic regions with less TEs density, giving rise to the 'two-speed genomes' concept (Plissonneau et al., 2017). The genome compartmentalization has been strongly associated with pathogenicity determinants diversification, that is because these regions are usually occupied by effectors which are exposed to mutation or duplication events mediated by the TEs around them (Dong, Raffaele & Kamoun, 2015).

An additional contribution of genomic and transcriptomic disciplines to the fungi study, is the elucidation of their functional richness. Databases are fed continuously by functional information from new fungal genome projects, which allows more accurate identifying the function of gene or protein sequence through homology-based methods. A clear example of that is the current knowledge of fungal pathogenicity which has been dramatically enhanced by the massive availability of pathogenicity related proteins/genes sequences that feed general and specialized databases (Fischer, 2003; Blin et al., 2013; Lombard et al, 2013; Jones et al., 2014; Nielsen, 2017; Urban et al., 2017; Rawlings et al, 2018). The systematic identification of proteins secreted during the disease development i,e., peptidases, lipases, CAZymes, SMs, and effectors, is a frequent task in genome assembly projects (Zhang et al., 2018; Lelwala et al 2019, Lopez et al., 2018). The collection of these proteins is known as “Secretome” and determining them in number and composition is elemental to predict fungal lifestyle as well as to identify the specific mechanisms driving their pathogenicity strategies.

Genomics advent also has contributed to clear up the phylogenetic relationships among fungal species. The fungal taxonomic replacement was initially inferred through morphological and biochemical traits (Zhang, Luo & Bhattacharya, 2017); however, these inferences were biased because some fungi structures cannot be propagated or identified in the laboratory, e.g., obligate biotrophic fungi (Stajich, 2017). Likewise, fungi taxonomic classification has been inferred through similarity analysis of a small number of conserved genes. Nevertheless, since each gene family exhibits a unique evolution rate and time scale, there is no selected gene taxonomical inference result that can entirely represent an organism (Choi & Kim, 2017). Conversely, whole-genome information can adequately represent an organism. In this sense, the availability of the complete sequence of orthologous genes and conserved protein families has allowed establishing more reliable phylogenetic relationships and with a better resolution (Zhang, Luo & Bhattacharya, 2017).

1.11. Generating a high-quality study resource, genome sequencing remarks.

Illumina platform is currently the most widely used sequencing technology because of its high yield, low cost, and error rate which is constantly optimized (Sato et al., 2019). This platform generates millions of short reads (up to 300 bp) with an error profile of less than 1% (Buermans and Dunnen, 2014). The current throughput of Illumina can grow up to 20 billion reads per line (Bedre, Avila & Mandadi, 2021). Typical applications of this technology include i) the assembly of draft genomes, ii) the discovery of genomic variants across populations, either from whole-genome re-sequencing (WGS) or from alternative cost-efficient protocols such as Genotype-by-Sequencing (GBS), iii) the identification of regulatory regions through the ChIP-Seq protocol, and iv) the estimation of gene expression levels at a given condition through the RNA-seq protocol (Trapnell et al. 2010).

The main limitation of Illumina sequencing technology is that the length of the read does not allow the proper identification of low complexity genomic regions (Koren and Philliphy 2015). It has been described that fungal genome assembly projects using short reads have been hampered by the high repetitive content and gene duplication events that affect the performance of the assembly algorithms (Compeau, Pevzner & Tesler, 2011). A classic example of this issue is the assembly of the genes coding effector proteins, as they present several copies that are usually impossible to decouple in Illumina-based assemblies (Dong et al., 2015). This was demonstrated by the AvrStb6 effector gene of *Zymoseptoria tritici*. The AvrStb6 effector was missing from the reference genome annotations for almost a decade (Goodwin et al, 2011), before mapping approaches allowed its identification (Zhong et al, 2017).

To tackle this issue with most of the current reference genomes, new sequencing approaches such as Pacific Bioscience (PacBio) and Oxford Nanopore aim to match Illumina's throughput and cost but offering a significant increase in read

length. Both perform direct sequencing of single DNA molecules with a length ranging from 10-20 Kbp (Eid et al., 2009; Niedringhaus et al., 2011; Mikheyev & Tin, 2014). These technologies produce a significant improvement in the assembly of genomes with a high percentage of repetitive regions because long reads can completely span copies of gene duplications and transposable elements (Schuster, 2007).

PacBio technology utilizes a DNA polymerase attached to supporting nanostructures known as "waveguides a zero point" (ZMW), which efficiently captures fluorescence signals in the real-time DNA fragments amplification (Buermans and Dunnen, 2014, Reuter et al., 2015). Instead of that, nanopore implements hundreds of independent proteic pores, adapted to an artificial surface, this single-molecule strategy measures the changes in the current generated by each nucleotide once DNA is threaded and subsequently passes through the nanopore (Reuter, Spacek & Snyder, 2015).

The main limitation of current long-read sequencing is the high error estimated at ~15% for both technologies, the long sequence errors are constituted by insertions, deletions, substitutions, and homopolymeric sequences (Laehnemann, Borkhardt & McHardy, 2016; Dohm et al.,2020). Hence, several alternatives to improve the error rate of long reads sequencing technologies have arisen. These alternatives include: the optimization of library preparation and sequencing protocols, ii) implementing hybrid sequencing strategies taking advantage of the low error rate of illumina reads in the correct long reads error phase (Alic et al.,2016), and improvements at algorithmic level (Zhang, Jain & Aluru, 2020). Currently, PacBio technology increased the read accuracy in 99.8%, using HIFI technology, a type of data that uses several sequencing passes generating Circular Consensus Sequences (CCS) which finally achieve an more accurate consensus (Wenger et al., 2019).

1.12. From the chaos to the order, solving the genome assembly problem

The development of genome assembly algorithms has followed the pace of the new developments in sequencing instruments and chemistries. The first-generation sequencing technology was represented by the Sanger method which produces reads with approximately 1’000 bps through a laborious methodology that incurred high costs. The Overlap Layout Consensus (OLC) algorithm was initially designed for the assembly of classical Sanger reads, solving the genome problem efficiently due to its flexibility with respect to read length and sequencing errors (Li et al., 2011).

The OLC algorithm has three main phases: i) The overlap phase performs pairwise alignments between reads identifying overlaps between them, next builds an overlap graph in which reads are represented by vertices (circles) and the overlaps between the reads are represented by edges. (blue arrows) (Figure 1-5). ii) The layout phase finds a Hamiltonian path over the graph built in the previous step (Figure 1-5), and, iii) The consensus phase builds a calculated consensus sequence aligning reads according to the layout identified in the second phase (Figure 1-5) (Miller & Koren, 2010). The main drawback of OLC algorithm is the computational cost, that increases with the number of reads, which is significatively high due to the all-against-all pairwise read alignments performed in the first phase of the algorithm (Li et al., 2011).

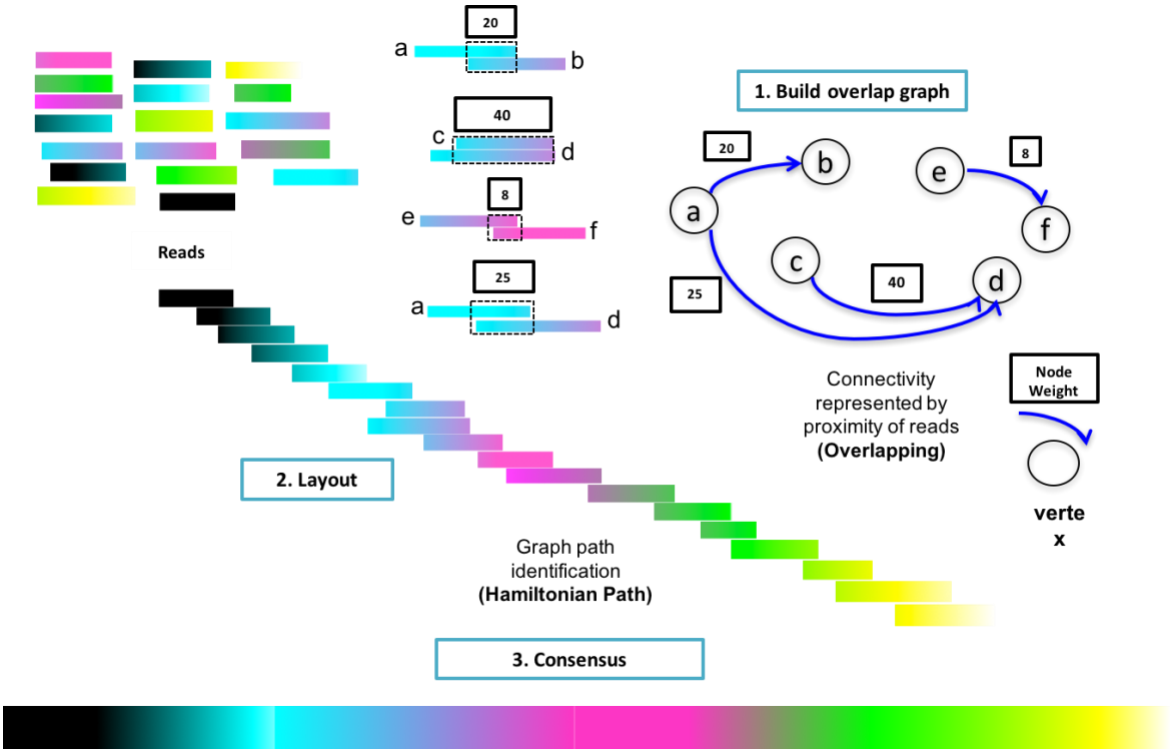


Figure 1-5. Graphic representation of Overlap Layout Consensus algorithm

Later second generation sequencing technologies represented by 454 Roche, Ion Torrent and Illumina platforms were developed. From these, the Illumina platform standed out over other technologies because of its low price, high yield and low error profile (less than 1%) (Buermans and Dunnen, 2014). The primary drawback of the Illumina platform is the short read length which, especially for larger and repetitive genomes, leads to highly fragmented assemblies (Koren and Philliphy 2015). Nonetheless, Illumina technology is currently the most widely used sequencing technology for *de-novo* genome assembly and re-sequencing projects.

As mentioned above, OLC had a high cost with increasing number of reads, with Illumina technology the number of reads for a given genome can increase by several orders of magnitude, making OLC not applicable. So, in order to deal with this new genome assembler tailored to large numbers of short read data implemented an alternative data structure called *de-Bruijn* graph for the consensus reconstruction. To that end, corrected or raw reads are chopped into substrings with

k length (*kmers*), which will be assigned to a directed edge that connects A and B nodes in the graph (Figure 1-6, black arrows). The nodes A and B, are defined by the *Kmers* prefixes (first *Kmer* -1) and suffixes (last *Kmer* -1) respectively (Zimin et al., 2013). Thus, a couple of nodes will be connected by a *kmer* if they have an overlap relationship, namely if they are neighbor *Kmers* (figure 1-6, in green). The draft genome consensus is then extracted looking for a Eulerian path, in which each edge will be visited only once and its sequence recorded only once respecting their direction (Figure 1-6).

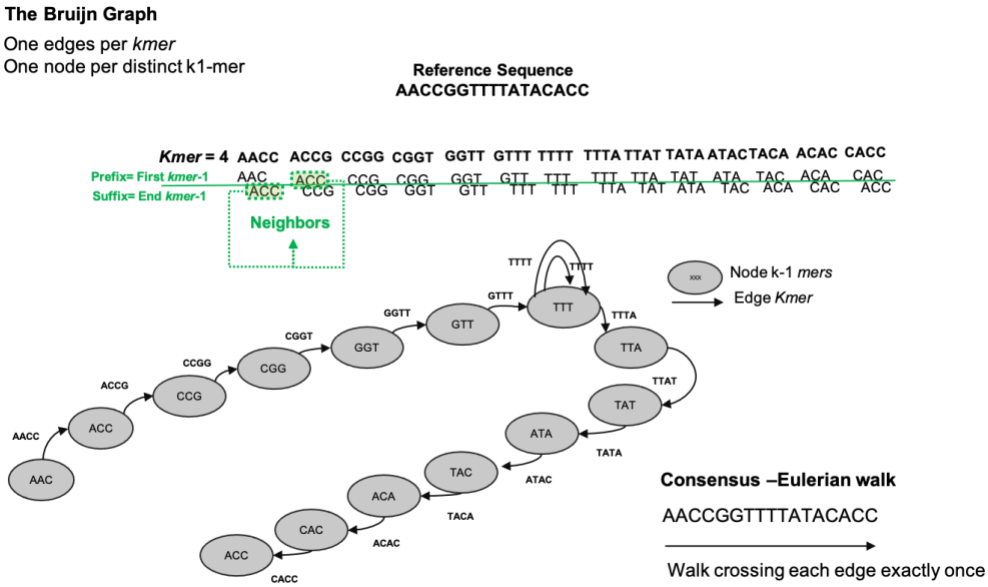


Figure 1.6. Graphic representation of the *de-Bruijn graph* construction for a genome assembly.

Representing the problem using *de-bruijn* graphs, do not require to compute expensive pairwise sequence alignments (Li et al., 2011). In turn, the disadvantages are the low tolerance to sequencing errors and genome structural properties like size, repetitive content, gene duplications, and heterozygosity, all of which can increase the graph complexity (Lin et al., 2016). Consequently, some short read assemblers implement sequencing error correction steps before the graph creation. Likewise, these programs increase the consensus accuracy by adapting graph correction steps to eliminate erroneous nodes, e.g., tips removal,

small repeats solution, and bubbles merging generated by repetitive elements (Li et al., 2009; Li et al., 2011). However, this approach is not effective to resolve different types of repetitive regions, even with long reads.

Third generation sequencing technologies represented by Pacific BioScience and Oxford Nanopore Technologies, generate longer reads (one to three orders of magnitude greater than second generation), allowing the assembly of more complex genomes, and a higher standard than previously possible with just second generation data. The assemblers for PacBio and Nanopore data reimplemented OLC as the base algorithm for the assembly process (Koren and Phillipy, 2015). To improve the computational efficiency of the process, and to take into account the high error rate of these new technologies, a new algorithm based on *kmer* information content and known as MinHash Alignment Process (MHAP) was implemented in the Canu assembler for the efficient identification of overlaps in noisy reads (Koren et al., 2017).

Canu algorithm runs in three phases, correction, trimming, and assembly. The correction phase improves the data quality performing multiple rounds of the read-to-read alignments prior to the assembly. Read overlaps are identified using MHAP, an adaptive *kmer* weighting scheme that identifies shared *kmers* between reads by comparing compressed reads sketches instead of individual *kmers* (Berlin et al., 2015). The main advantage of MHAP usage in the correction phase is that the number of *kmer* occurrences is informative, thus repetitive *kmers* can receive low weights, which allows identifying and saving this information for the algorithm in further steps (Assembly) (Koren et al., 2017). Other assembly strategies often exclude repetitive reads from the original dataset generating shorter and more fragmented genomes (Chin et al. 2016).

The read correction phase replaces the original noisy read sequences with consensus sequences computed from overlapping reads which are saved in a database to be used as an input in further processes. The trimming phase identifies

and eliminates unsupported regions, sequencing adapters, and chimeric sequences in the input. The consensus sequence is obtained constructing the Best Overlap Graph (BOG), from which only overlaps that overcome an error rate threshold (automatically defined by Canu software based on the sequencing data error), are saved in a new overlaps database. The longest overlaps are recomputed using only the filtered subset and the consensus prediction is performed from the graph following the maximal non branching paths (Koren et al., 2017). The final step of the Canu assembly process is known as polishing. In this step, reads are realigned to the genome assembly generated by BOG and base pair errors are corrected based on consensus. The Illumina reads can also be aligned at this step to further improve the correction process taking advantage of their lower error rate (Koren et al., 2017).

Second and third sequencing generation data could be integrated in hybrid assembly approaches, in which a draft genome is constructed from long reads using any of the assembly algorithms already described. Subsequently, Illumina reads are often aligned against the consensus in the polishing process (Miller et al., 2017). Alignments of Illumina reads could also be used to identify potential events of local duplications or even chromosome aneuploidies (Gallone et al., 2016).

1.13. Looking for the genome function, structural and functional annotation

Once a genome sequence is assembled, the usual next step is to identify different functional elements such as gene and gene families, regulatory regions, and transposable elements in the assembled sequence. This process is called structural annotation of the genome sequence, usually shortened to genome annotation (Mukund et al., 2001, Muñoz et al., 2012) and could be performed using *ab-initio* or homology-based approaches depending on the available biological evidence (Hoff & Stanke. 2015). The *ab-initio* methods predict the gene models performing statistical analysis using just the genome sequence, exploiting specific features like codon frequencies and intron-exon length distribution (Ejigu & Jung. 2020; Yandell

& Ence. 2012). This analysis effectively predicts protein-coding sequences (CDSs) but is not efficient enough to predict untranslated regions (UTRs) and splicing alternatives (Yandell & Ence. 2012). Homology based methods are more accurate in the gene prediction since it is based on alignments of biological evidence against the annotated genome. Protein sequences expressed sequence tag (ESTs) and transcript sequences (RNAseq) are the primary biological source used to train the gene predictors in homology-based methods (Keilwagen et al,. 2018).

Protein spliced alignments contribute with exon structure and translation evidence. Due to the amino-acids conservation level between species, this resource is not restricted to be authentic to the annotated organism, and closest related protein families can be used as a complementary resource (Yandell & Ence. 2012). On the other hand, RNAseq data has a massive potential to improve the gene predictors' sensitivity, and it provides location and information about the structure of transcripts (Keilwagen et al,. 2018). Transcript alignments are performed with splicing-aware programs like GeneMark-ET (Lomsadze et al., 2014), providing exon boundaries evidence to the gene annotation. Read coverage is used to indicate intron positions which strengthen the identification of splice sites (Hoff & Stanke. 2015).

Once the genes are structurally predicted from the genome assembly, the next step is to identify the specific functions associated with these protein or nucleotide sequences. This systematic process is usually called functional annotation and often implements homology searches coupled to probabilistic models to associate a function to a given sequence (Sivashankari & Shanmughavel, 2006). The most common tools to functionally annotate a proteome are BLAST suite programs or HMMER3 software suite, this latest implements Hidden Model Markow, calculating likelihood scores and expectation values (E-value) per each aligned protein residue (Eddy, 2009) . These algorithms compare the predicted proteins against widely used databases, like Pfam (Punta et al., 2012), TIGRFAMs (Haft et al., 2012), and SMART (Letunic et al., 2012) between others. The functional annotation can also be complemented by other approaches such as machine learning and data mining

and cluster analysis studying the relationships between genomes of different species identifying proteins that are expected to evolve from the same metabolic pathway or a structural complex shared among the compared species (Sivashankari & Shanmughavel, 2006).

Chapter 2

2. The *P. ulei* genome structure

2.1. Introduction

Fungi are ubiquitous eukaryotic and heterotrophic microorganisms that establish biological relationships with other living organisms to carry out their nutrition processes and complete their life cycle (Blackwell, 2011). Taxonomic reports in the 90s described that the fungi kingdom harbors approximately 100,000 species matching in number that of the animal kingdom and surpassing those in the plant kingdom by a 6:1 ratio (Webster & Weber 2007). However, it was only at the beginning of the 21st century, with the development of culture-independent methods and high-throughput sequencing technologies, that this group's richness was appreciated. Most fungi are non-pathogenic, such as arbuscular mycorrhizae and endophytic fungi (Petrini, 1991; Arnold, 2003; Deshmukh & Rai, 2005). Only 10% of them are pathogens and affect mainly vegetal hosts (Knogge, 1998). Despite this low representativity, phytopathogen fungi generate an economic impact estimated at 200,000 million dollars annually, constituting the greatest threat to agricultural and commercial crops (Horbachab et al., 2011) and constant risk for food security and the global economy (Mooibroek, 2000). Therefore, the study of phytopathogen fungi is crucial from the phytosanitary point of view since the development of new control strategies could be plausible through a holistic understanding of their pathogenicity.

The ascomycete fungus *P. ulei* is the causal agent of the South American Leaf Blight (SALB), a disease that represents the main threat for Latin-American natural rubber plantations (Chee, 1976). SALB affects the photosynthetic functions of the rubber tree by increasing defoliation, thus reducing its production by up to 75% (Chee & Holliday, 1986). Despite the *P. ulei* agronomic importance, the knowledge around the biological and pathogenic behavior of this pathogen is minimal. The previous has been strongly associated with its presumed obligate biotrophic lifestyle which

hinders its *in-vitro* growth in absence of the host. Besides, the lack of a reference genome for *P. ulei* has been the major constraint to determine the mechanisms of pathogenicity involved in SALB.

Nowadays, the assembly of a reference genome for a particular species is a building block for most primary and applied research. The development of high-throughput sequencing technologies followed by reductions in sequencing costs (Goodwin et al., 2016) enabled extensive surveys of fungal genomes, converting them into the most sequenced eukaryotic group (Jun-Ma and Fedorova 2010; Galagan et al., 2005). The accessibility to this information supported by advances in genomics, and comparative studies, have revealed a wide variability in fungal genome architecture (Stukenbrock & Croll, 2014; Möller and Stukenbrock, 2017; Mohanta & Bae, 2015), exposing part of the hidden complexity in this group. Despite the numerous contributions to the knowledge of fungal biology gained in recent times, there are still many gray and dark areas that require further study.

The fungal genome size varies from <17 Mbps to >177 Mbps. This variation has been directly correlated with the content of transposable element (Haridas et al., 2020). The majority of fungal genomes are occupied by 1-25% of repetitive elements (Rao et al., 2018). However, some fungal phytopathogens exhibit exceptional transposable elements percentages, e,g., *Blumeria graminis formae speciales tritici* which contains 90% of transposable elements (Wicker et al., 2013). The repetitive elements are the primary drivers of fungal genome plasticity and evolution, acting as mediators of adaptation to biotic and abiotic stress (Stajich, 2017). The fungal genomes' gene content goes from a few 10 thousand to a few 20 thousand, being 11'129 the average in Ascomycota species (Mohanta & Bae, 2015). This variation has been associated with the nutritional lifestyle, where the degree of dependence on the host could lead to a gene content reduction (Hane et al., 2020).

Furthermore, genomic and comparative studies have positively impacted the fungal taxonomy field, increasing the identification of new species (about 1200 annually

just during the last decade) with a better resolution than the classical taxonomic studies. The taxonomic relationships among fungi species were initially defined through similarity analysis of a small number of conserved genes (multi loci analysis). However, these inferences could be biased given that each gene family exhibits a unique evolutionary rate and the analysis of few genes cannot entirely represent an organism (Choi et al., 2017). Genome sequencing projects allow the determination of the orthologous genes content, offering an opportunity to perform a better resolution phylogenetic classification for a species (Emms & Kelly, 2015).

This chapter will outline the *P. ulei* genome sequencing, assembly, and structural annotation strategies that lead to obtaining a high-quality genome proposed in the first objective. We will describe the *P. ulei* genome architecture in terms of ploidy, gene density, and transposable element content. Also, considering that the *P. ulei* gene catalog is a source obtained from the genome structural annotation described in this chapter, we will include the *P. ulei* phylogenetic classification in the *Pseudocercospora* genus and the Mycosphaerellaceae family.

2.2. Material and methods

2.2.1. Sampling and Fungal isolation

P. ulei GCL012 was obtained from the *H. brasiliensis* GT1 accession established at the Agrosavia clonal garden in Villavicencio - la Libertad, Colombia. The source for DNA and RNA extraction were stromas of the isolated *P. ulei* GCL012 (Figure 2-1). *P ulei* GCL012 was isolated from conidia collected from a single plant lesion (Figure 2-1 a) and grown on M3 solid medium (Gasparotto, 1997) at 25ºC in the dark for 45 days (Figure 2-1, blue box) until the stroma formation was visible. The fungal biomass was propagated by macerating the obtained stromata in 2 mL microcentrifuge tubes (Eppendorf®, Germany) and then transferring to 125 mL flasks with M4 sporulation solid medium (Figure 2-1 e-h). The M4 media contains potato broth, amino acids, and peptone (Gasparotto, 1997). The sporulation was stimulated by exposure to white light for 90 minutes/day (Lieberei, 2007).

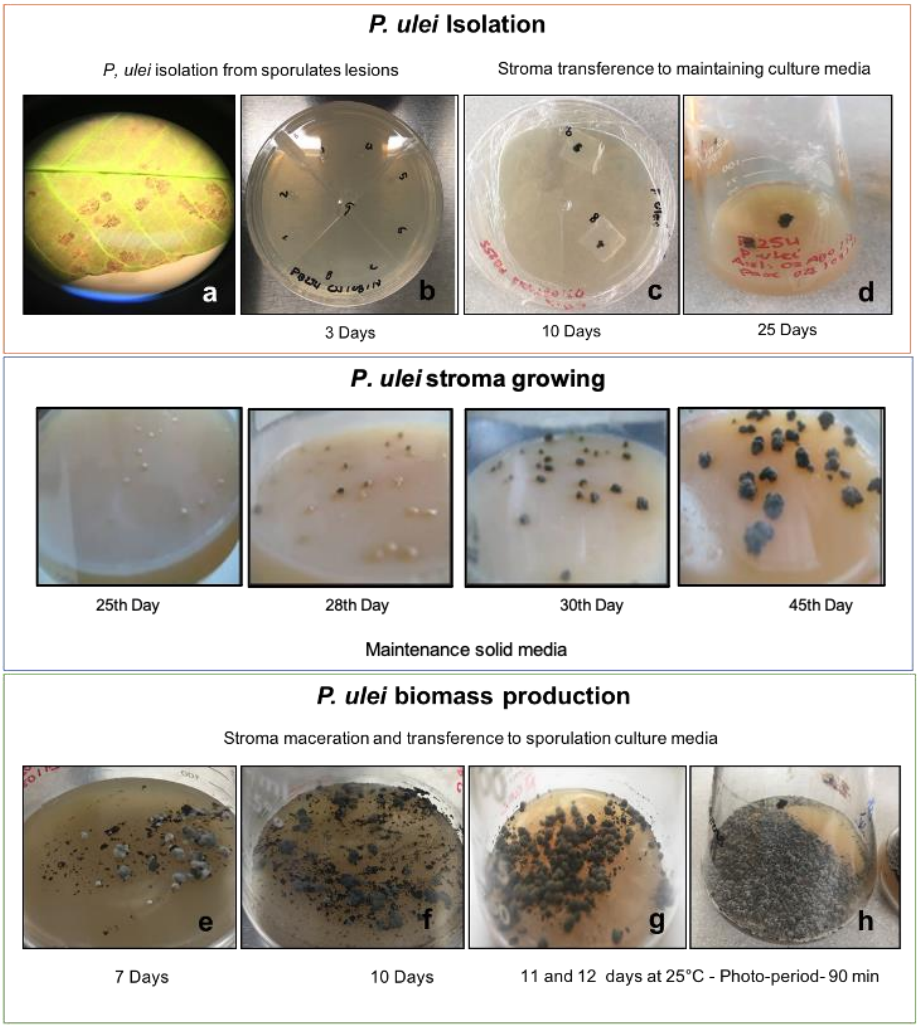


Figure 2-1. Graphic representation of *P. ulei* isolation and *in vitro* culture.

2.2.2. Microscopic and molecular identification of *P. ulei* isolates

The *P. ulei* isolates were identified by microscopy and molecular approaches. Microscopic identification of *P. ulei* conidia, mycelium, and stroma morphology were confirmed from *in vitro* culture isolates. To that end, the fungal structures were placed in microscope slides and stained with lactophenol blue solution. The observations were performed in an Olympus Microscopy CH30 model, serial: T39D15816. The macroscopic observations were performed in the Advanced optical stereoscopy, serial:0709011.

For the molecular identification we amplified both TDF-Mu1 (Transcription-derived fragment – *P. ulei,* a hypothetical apoplastic effector previously identified) and ITS genes. The ITS region was amplified using 20µM of ITS-1 and ITS-4 primers (Table 1-1). The PCR reaction consisted of 50µM of MgCl2, 20µM of dNTPs, 10X of reaction buffer, 5 U of Taq polymerase, and 25 ng/µl of *P. ulei* DNA. The final volume was completed to 50 µl with molecular grade water. The reactions were incubated at 95ºC for two minutes followed by 30 cycles of denaturation 95ºC for 60 seconds, annealing 50ºC for 60 seconds, and elongation at 72ºC for 60 seconds. The amplification protocol finished with a seven minutes elongation step of 72ºC.

The TDF-Mu1 PCR reaction was performed in a total volume of 25 µL consisting of 2.5 µL of 10X Buffer, one µL of MgCl 2 (50 mM), 0.3 µL of dNTPs (20 mM), 1.25 µL of primers at a concentration of 10 µM (Table 1-1), 2.5 U of Taq DNA polymerase Fermentas® and five µL of DNA (5ng / µL). The PCR amplification conditions consisted of: an initial temperature of 94ºC for 5 min, followed by 34 cycles with the times and temperatures of 94ºC for 30 sec, and 57ºC for 30 sec, and 72 ºC for 30 sec, finally, a 10 min elongation step was performed at 72ºC. The amplified fragments' size was confirmed by electrophoresis in agarose gel at 1.5% compared with the hyper ladder 25bp marker Bioline®.

Table 1-1. Primers sequences for *P. ulei* isolates identification.

Primers	Sequences	Reference
TDF2R	5'- AGAGCGGTCGGTTGACGGTA -3'	Garcia, I, 2012
TDF2F	5'- TGCCACCTACGTTCCTGTCCA -3'	
ITS1	5'- TCCGTAGGTGAACCTGCGG -3'	White et al.,1990
ITS 4	5'- TCCTCCGCTTATTGATATGC -3'	

2.2.3. Genomic DNA extraction

High molecular weight (HMW) of total DNA from sporulated stromata was extracted following Stirling's protocol (2014), modified by adding phenolic extraction followed by three phases of chloroform extractions. DNA concentration, integrity, and purity were assessed by fluorometry (Qubit®), agarose gel electrophoresis (1%) with Tris-borate-EDTA (TBE) buffer stained with SYBR safe (0.5 mg/L), and spectrophotometry (Nanodrop®), respectively.

The identity of the extracted DNA was confirmed by sequencing the ITS region by sanger technology. The ITS PCR product was purified using UltraClean® PCR Clean-Up Kit and sequenced by Sanger in the Genetics institute from the Universidad Nacional de Colombia. Finally, we performed local alignments using the BLASTn algorithm against the NCBI public database.

2.2.4. Genomic DNA sequencing

P. ulei HMW total DNA was sequenced by a hybrid strategy producing second and third generation sequencing reads. Third generation sequencing reads were produced by the Pacific Bioscience (PacBio) Pacific BioSciences and Oxford Nanopore Technologies (ONT) platforms, and second-generation sequencing reads were generated by the Illumina platform. PacBio data set consisted of 10 Gb of raw data which was generated on a Sequel instrument by Novogene (USA) using a 1M SMRTBell. For ONT, a total of 8 Gb of raw data was generated running the *P .ulei* library for 36 hours in the MinION device using a single R9.4.1 flow cell. The library construction was performed with the 1D Genomic DNA by Ligation kit (SQK-LSK109), modifying the manufacturer conditions in the DNA elution steps increasing the temperature and incubation time at 37ºC and 30 min, respectively. Base-calling was performed with the Guppy base-caller software.

Illumina consisted in 10 Gb of paired-end reads (2x250 bp) which were generated on a HiSeq 2500 instrument by Novogene (USA) using the TruSeq library preparation kit. Short reads were trimmed with Trimmomatic v.0.38 (Phred score-

33, leading-3, trailing-6, sliding-window-4:20, milen:80) (Boger, 2014). Contaminant sequences were identified by BlobTools version 1.0.1 (Laetsch and Blaxter, 2017) and subsequently eliminated by mapping from the illumina dataset using BBduk2 v. 35.85 (Bushnell, 2014). For that purpose, the contaminants genomes, as well as the *P. mori* mitochondrial genome were downloaded from NCBI (See Table 1-2 for a list of genome IDs) and then mapped against illumina raw data, keeping only the unmapped reads for downstream processes.

Table 1-2. contaminant genomes IDs identified in Illumina data

Contaminat Genomes	IDs
Arthrobacter alpinus ERGS4:06	ASM144557v1
Arthrobacter castelli DSM 16402	ASM43070v1
Arthrobacter enclensis	ASM145702v1
Arthrobacter globiformis	ASM23891v2
Arthrobacter crystallopoites	Arthrobacter crystallopoitesv3.1
Arthrobacter alpinus R3.8	ASM129462v1
Cupriavidus basilensis	ASM83230v1
Cupriavidus gilardii	ASM128146v1
Cupriavidus metallidurans	ASM19601v1
Cupriavidus necator	ASM869338v1
Cupriavidus pauculus	ASM869338v1
Glutamicibacter halophytocola	ASM130256v1
Herbaspirillum seropedicae	ASM104094v1
Microbacterium laevaniformans	ASM158460v1
Nocardia farcinica	NCTC11134
Rubrobacter aplysinae	ASM102950v1
Rubrobacter radiotolerans	ASM66189v1

Rubrobacter xylanophilus	ASM1418v1
Sinorhizobium fredii	ASM28389v1
Zobellella denitrificans ZD1	ASM223736v1
Zobellella denitrificans F13-1	ASM240748v1
Pseudocercospora mori mitochondrion	NC_037198.1
Pseudocercospora mori mitochondrion	MG543071.1

2.2.5. *P. ulei* GenomeScope profile

The characteristics of genome structure represented by genome size, polyploidy, heterozygosity and repetitive content, are the primary hurdle in the consecution of a complete genome assembly, especially those projects that only include second generation sequencing data (Liu et al.,2013). The *kmer* spectrum represents an accurate approach to estimate genome architecture features (Vurture et al.,2017). We determined the overall characteristics of the *P. ulei* genome structure using GenomeScope v.2.0 giving as an input file the *kmer* counts obtained with Jellyfish v.2.2.10 (Marcais and Kingsford, 2011).

The diagram represented in Figure 2-8 describes the general methodology followed to obtain the *P. ulei* genome assembly and gene annotation; in grey dotted line boxes we show the steps, and purple boxes the programs implemented.

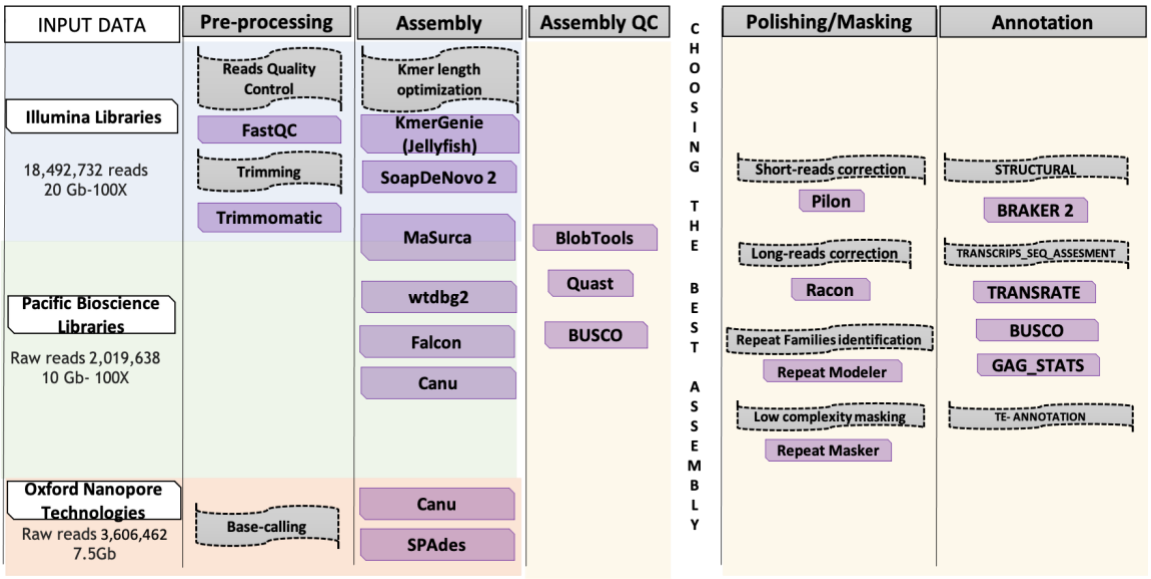


Figure 2-2. General methodology implemented for *P. ulei* genome assembly and gene annotation.

2.2.6. Genome assembly

2.2.7. Hybrid, non-hybrid, and hierarchical assembly methods were assessed

The *P. ulei* short reads assemblies were reconstructed with *de Bruijn graph* data structure using SoapDeNovo2 (Luo et al., 2014), setting five *kmer* values (61, 71, 81, 101, and 127). Given that the user decides the *kmer* length parameter in SoapDeNovo2 assembler, we used the kmergenie v.1.7051 to make an informed decision about the best value of k for the assembly. Kmergenie constructs abundance histograms for several *kmer* lengths finding the optimal *k* value (Chikhi & Medvedev, 2013). Furthermore, we assembled *P. ulei* short reads using the MaSuRCA (Maryland Super-Read Celera Assembler) software v. 3.2.9, an assembler that implements a different genome assembly strategy, reducing the paired end reads complexity by transforming them into lower coverage super-reads (Zimin et al., 2013).

Long-read genome assemblies were built using PacBio, and ONT data sets independently (non-hybrid approaches for assembly) and also giving both datasets as arguments in some assembly strategies (hybrid approaches). PacBio polymerase reads bam files were converted to fasta format with the bamtofastq from bedtools suite v. 3.3 (Stanke et al., 2010). The presence Circular Consensus Sequence (CCS) in the PacBio data was checked with the ccs tool from the pbccs software (https://github.com/nlhepler/pbccs). Single-technology assemblies (non-hybrid, i.e., only ONT or only PacBio) were generated with the wtdbg2 v. 2.3 (Ruan, 2019), PB-assembler, known as Falcon (Chin et al., 2013; Chin et al., 2016), and Canu (Koren, 2017) software packages. The hybrid assemblies were built with MaSuRCA v. 3.2.9 (Zimin et al., 2013) and SPAdes software v. 3.15.1 (Bankevich et al., 2012).

The quality metrics for each assembly were calculated with Quast v.5.0.2 (Gurevich et al,.2013). Assembly completeness was assessed by Benchmarking Universal Single-Copy Orthologous (BUSCO) software v.3.0.2 with the Ascomycete dataset from OrthoDB v.9. (Simão et al., 2015). The assembly with the best quality metrics and highest completeness was selected for further processing. The selected final assembly was polished by two rounds of Pilon v.1.23 (Walker *et al*,.2014) using the Illumina dataset.

2.2.8. Genome scaffolding

We exploited the ONT reads to scaffold the genome assembly using PyScaf v1.0 (https://github.com/lpryszcz/pyScaf) and LINKs v.1.8.6 (Warren, 2014) tools. PyScaf software aligned ONT against the *P. ulei* assembled contigs, identifying reads connecting two or more contigs and then joining the adjacent contigs (Figure 2-3 A), We assessed the performance of this program with raw and corrected ONT reads. To verify the merged contigs veracity, we mapped ONT reads against *P. ulei* genome with Minimap2 v.2.15 software (Heng, 2018), and then identified all reads that match two contigs with a manual R-script.

A second scaffolding strategy (Figure 2-3 B) consisted of identifying *kmer* pairs between ONT raw reads and the assembled contigs using the LINKs v.1.8.6 software. The best scaffolding strategy was chosen with respect to the efficiency of merging contigs into scaffolds without affecting the genome length and the completeness BUSCO profile.

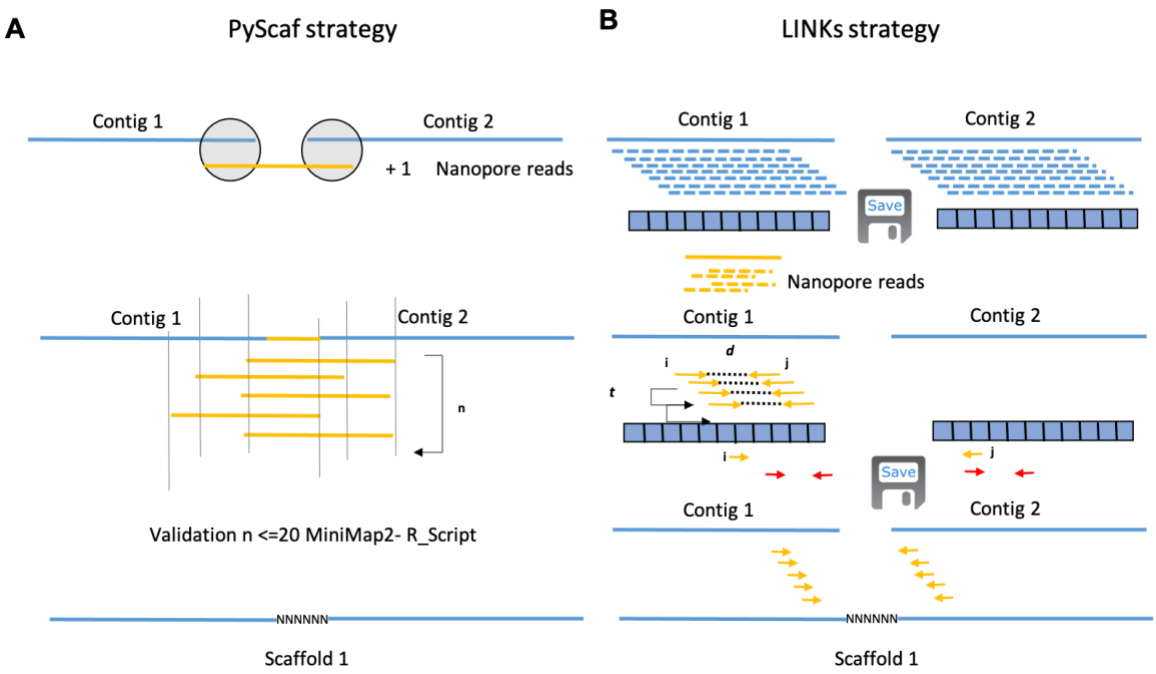


Figure 2-3. *P. ulei* Genome scaffolding approaches

2.2.9. Identification of repeats and masking

The *de novo* repeat families identification of the *P. ulei* scaffolded genome were performed using RepeatModeler v.1.0.11 (Smit, Hubley & Green, 2015). The consensus file generated as an output from RepeatModeler was used to mask the *P. ulei* scaffolded genome using RepeatMasker v.open-4.0.9 (Graovac & Chen, 2009). The softmasking conversions were performed with maskfasta tool from bedtools suite (Quinlan & Hall, 2010).

2.2.10. Structural genome annotation

2.2.10.1. *P. ulei* transcriptomic data generation

We have generated RNAseq data of *P. ulei* to support our gene prediction.

The growth stages of the fungi were used for RNA extraction, i.e., early sporulation (13th day of growth) and later growth (18th day of growth). Two RNA extraction protocols were assessed to achieve the RNA purity and concentration parameters, i) RNA standard protocol which performs a membrane solubilization using SDS, an

organic extraction with phenol-chloroform in 19:1 proportion and finalize with several cleaning steps only with chloroform (Kingston, 2000), and ii) Using the NucleoSpin® RNA Plant and Fungi MACHEREY-NAGEL kit (Düren, Germany), that consists in a total RNA extraction protocol using purification columns.

One library per growth was prepared using i5 and i7 index library preparation kit, mRNA was enriched using polyA primers. A total of 341'630'429 PE-150bp strand-specific raw RNAseq reads were produced in a Illuminas' NovaSeq 6000 instrument (San Diego, CA). The RNA was pre-processed with Trimmomatic software v. 0.38 (Bolger, 2014) with the following parameters SLIDINGWINDOW:4:20 and MINLEN:80. Putative contaminant sequences were identified with Kaiju tool v1.6.3 (Menzel 2016), rRNA sequences were identified by mapping them against *P. ulei* ribosomal sequences available in public databases. Putative contaminants and rRNA were eliminated from the dataset with BBDuk2 software (Bushnell, 2014).

The *Pseudocercospora spp* introns evidence was generated mapping both *P. ulei* RNAseq reads and RNAseq reads from *P. fijiensis* (SRP075820) against the *P. ulei* genome using HISAT2 v.2.0.5 (Kim & Salzburg, 2015) and converted into hints using bam2-hints v. 3.3 (Stanke et al., 2006).

2.2.10.2. Gene prediction optimization

Genome annotation was performed with Braker2 v.2.1 (Hoff et al., 2015) training predictors with identified BUSCO genes, protein sequences from *Pseudocercospora spp*., and *P. ulei* and *P. fijiensis* introns evidence described in 1.1.9.1 paragraph.

Gene prediction optimization was performed by combining evidence sources in seven annotation pipelines. The first six pipelines consisted of the masked genome assembly used: 1) The protein catalog of *Pseudocercospora fijiensis*, 2) a non-redundant database of *Pseudocercospora* sp. proteins (p_NRDB), 3) a set of high quality exon-exon borders (filtered hints) from *P. fijiensis'* RNAseq, 4) the set of

identified BUSCO genes, 5) a non-redundant database of *Zymoseptoria tritici* and *Passalora fulva* proteins (m_NRDB), 6) *P. ulei* gene models.

The seventh annotation pipeline exploited the unmasked genome, concatenating all the above-mentioned evidence sources. The quality of the structural annotation results was evaluated, determining the orthologous genes composition with BUSCO software (Simão et al., 2015), and obtaining the quality transcripts metrics with GAG statistics (Hall & Geib, 2014). The quality of the annotation was also assessed through OrthoFinder v.2.3.3 (Emms & Kelly, 2019) determining the gene percentage assigned to orthogroups including the protein catalogues from *P. eumusae, P. fijiensis, P, musae* (Chang, Salvucci & Crous. 2016) *P, macadamia* (Akinsanmi & Carvalhals, 2020), and *P. pini*, *P, cruenta, and P, fuligena* (as their annotation was not available, we generated them as part of this project).

Finally, the best gene models were selected with the Evidence Modeler (EVM) program v.1.1.1 (Haas et al., 2008) using the *ab initio* gene predictions and transcripts alignments evidence. To select the most informative *ab initio* predictions, we identified the unique single and duplicated orthologous genes assigned in every single annotation pipeline choosing those that contribute with unidentified orthologous.

Transcript alignment evidence was generated assembling the *P. ulei* transcriptome by *de novo* and reference guided approaches using Trinity v.2.8.5 (Haas et al., 2013) and StringTie v. 1.3.6 (Pertea et al., 2015; Pertea et al., 2016) programs, respectively. Candidate coding regions were identified aligning the *de novo* assembled transcripts against *P. ulei* scaffolded genome with Transdecoder program v. 5.5.0 (https://github.com/TransDecoder/TransDecoder/wiki). Each of these lines of evidence was presented to EVM in separated tracks, assigning the weights to each evidence as follows: Braker6 = 2, Braker_7 = 6, Augustus_genome = 10 and Transcripts alignments = 10.

2.2.11. *P. ulei* phylogenetic replacement

In order to place *P. ulei* GCL012 among the *Pseudocercospora* genus, we performed a phylogenetic analysis based on the four genomic loci: actin, partial transcription elongation Factor 1-α (EF-1α) and ITS 1 and 2, including 278 *Pseudocercospora* species, whose data was publicly available (See Appendix 1 for the accession number for each marker in each species). The four genomic loci dataset were extracted from the GCL012 genome by EMBOSS:extractseq program, then aligned with MAFFT v 7.407 (Katoh et al 2005) using 'auto' alignment option for the Actin and EF-1α loci and 'mafft-qinsi' for the ITS 1 and 2 loci. Poorly aligned regions were identified and removed with TrimAl v 1.4rev22 using the -automated1 mode (Capella et al., 2009) and Jalview Software v 2.11.1.3 (Waterhouse et al,. 2009). We carried out a partitioned phylogenetic inference with IQTree v1.6.9, using a partition matrix generated with FASconCAT-G v 1.04 (Kück & Longo. 2014). The evolutionary model for each loci were: used for the phylogenetic inference were Actin:TIM2e+G4, EF-1α:TN+F+G4,TPM2+F+R3, and ITS1-ITS2:TPM2+F+G4.

In addition, we carried out a phylogenetic inference using 1315 orthologous gene sequences extracted from the genome sequences of seven species within *Pseudocercospora*, and 37 species outside *Pseudocercospora* genus (See Appendix 2, for accession numbers). Amino acid sequences from orthologous genes shared between the 44 assessed species were independently aligned with MAFFT v7.407 (Katoh et al 2005), and trimmed by trimAL v_1.4rev22 (Capella et al., 2009). Maximum likelihood phylogeny was estimated with IQTree v1.6.9 using 1000 bootstrap replicates. All tree files inferred were concatenated with Astral v.5.7.3 (Zhang et al., 2018).

2.3. RESULTS AND DISCUSSION

2.3.1. Microscopic identification of *P. ulei* isolates

The morphological identification of *P. ulei* GCL012 strain presented at a macroscopic level the presence of irregular shaped black stroma with short olivaceous-brown mycelia with velvety appearance during the sporulation phase (Figure 2-4 a). At the microscopic level, we observed stromatic hyphal aggregations (Figure 2-4 b). The conidia and mycelia morphology matched with the *P. ulei* anamorphic phase *Fusicladium heveae* described by (Schubert, Ritschel & Braun, 2003) that reported that the asexual conidia are obclavate, straight to slightly curved and septated as is shown in the (Figure 2-4 d). The collected stroma is maintained in the molecular characterization laboratory's culture collection in the biotechnology institute from the National University of Colombia (IBUN).

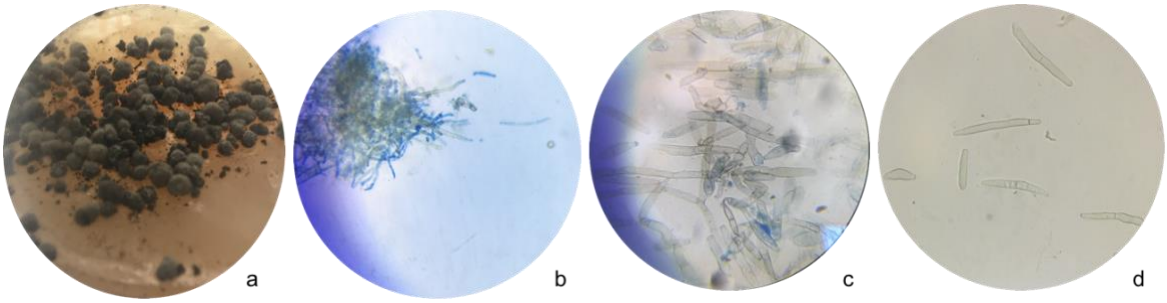


Figure 2-4: Macro and Microscopic structures of *P. ulei in vitro* culture. a) Stereoscopy observation at magnification 1X of growth of stroma on a Erlenmeyer 125 mL with sporulation artificial media (M4) at 12 days, b) Mycelia of *P. ulei* (450X) lactophenol blue solution stain, c) Conidia formation (1000X) lactophenol blue solution stain, and d) *P. ulei* conidia obtained *in vitro* culture. b, c and d photos were observed in a Olympus Microscopy CH30 model and taken with a iPhone 6s camera.

2.3.2. *P. ulei* GCL012 DNA extraction standardization

The DNA extraction of the *P. ulei* GCL012 isolate was performed following the Stirling protocol (2003) with some modifications. The DNA extraction standardization consisted of assessing several lysis processes to guarantee good

DNA integrity, compatible with third generation sequencing technologies. Liquid nitrogen and zirconia beat lysis options were assessed from 0.5 g of fungal biomass. Liquid nitrogen lysis showed better results in terms of DNA integrity (Figure 2-5 a - lines 1 and 2); however, the concentration was less than the obtained with Zirconia beads in all the cases (Table 1-3, samples DiPu166a -b, D2Pu167a -b, and D3Pu68a -b). Despite the good concentration values obtained with Zirconia beads, this lysis method was highly detrimental for DNA integrity (Figure 2-5 a - lines 3, 4, and 5), even reducing maceration time.

We achieved the DNA concentration for the long reads sequencing platforms, increasing the biomass quantity to 1 g, and macerating it with liquid nitrogen three times in small mortar until all fungal biomass became powder (Figure 2-5 b lines 1-8, corresponding to samples T1-T4 and BMK1- BMK4). However, by increasing the starting fungal biomass, the purity of the extracted DNA as defined by the relationship between the absorbance ratios 260/230 and 260/280 was less than 1.3, which required further processing. To improve the DNA purity, we added two organic extraction phases (chloroformic) to *P. ulei* DNA extraction protocol. This improved the absorbance relationship values (Table 1-3, samples T1-T4 and BMK1-BMK4), which were determinant to achieve the long reads platforms' DNA quality requirements. Finally, DNA integrity results assessed in agarose gel by comparison with the concatemer λ DNA marker, shown that the band integrity of the liquid nitrogen macerated samples (Figure 2-5 b lines 1-8 *P. ulei* DNA) presented a bigger weight than the λ DNA marker, which has 48’502 bps (Figure 2-5 b lines 9-11: corresponding to 50, 100, and 150 ng/µl λ phage).

Once the DNA extraction enables to achieve the DNA quality parameters for PacBio and Illumina platforms, the samples with the best concentration, integrity and purity parameters were sent to the sequencing facility for raw data production.

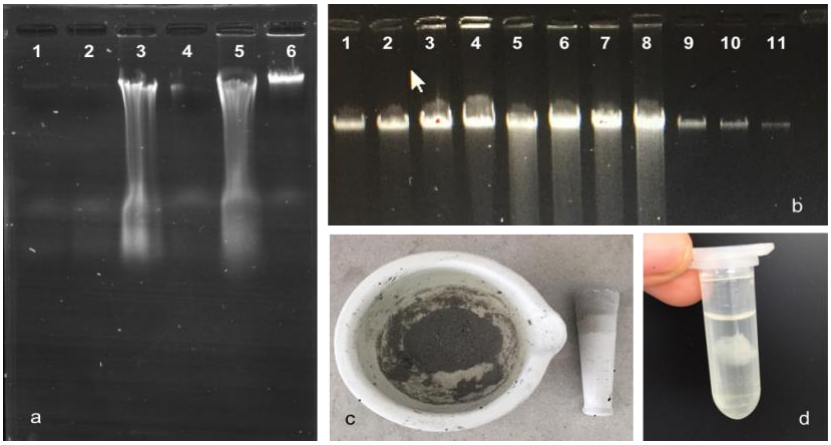


Figure 2-5. Electrophoresis in agarose gel 1.0%, a) Lysis protocols assessment: Liquid nitrogen lysis corresponds to lines 1 and 2 and Zirconia beads corresponds to lines 3, 4, and 5. b) DNA integrity obtained duplicating the fungal material to start with: lines 1-8 correspond to DNA extractions from *P. ulei* GCL012 isolated named T1-T4 and BMK1-BMK4. c and d) fungal biomass pulverization and DNA precipitation photos taken with a iPhone 6s camera.

Table 1- 3: Quality parameters of DNA extractions from *P. ulei* GCL012 isolated.

Method	ID Sample	ng/µl	µg totales	260/280	260/230
Zirconia beads	D1Pu166a	488	24	1.94	1.4
	D1Pu166b	292	14.6	1.95	1.3
	D2Pu167a	520	26	1.8	1.46
	D2Pu167b	320	16	1.90	1.4
	D3Pu68a	393	19	1.93	1.3
	D3Pu68b	202	10.1	1.7	0.9
Liquid Nitrogen	T1	80	8	2	1.66
	T2	50	5	2.1	2.2
	T3	268	26.8	2.15	2.2
	T4	46	4.6	2.21	2.23
	BMK 1	82	8.2	2.14	1.96
	BMK 2	127	12.7	1.98	1.4
	BMK 3	47	4.7	2.14	2.15
	BMK 4	91	9.1	1.18	2.28

The DNA integrity and purity are the more detrimental variables to the downstream analysis of sequencing data (Jun-Ma & fedorova, 2010). It was demonstrated by Nietsch and collaborators, who described that besides affecting the fragment size during the sequencing process, these variables, interfered with the number of detected variants in human DNA samples with low quality (Nietsch et al., 2016). The

P. ulei stroma rigidity was the main obstacle to achieve suitable DNA integrity and concentration for long reads production. We prioritized the DNA integrity in our lysis standardization process to avoid bias in the DNA concentration measures generated by its fragmentation (Sedlackova et al., 2013) and to optimize the size of long reads.

2.3.3. Molecular confirmation of *P. ulei* DNA identity

The identity of the GCL012 *P. ulei* strain was confirmed by sequencing the ITS and TDF-Mu1 regions. Agarose electrophoresis results showed the presence of the ITS's amplimer constituted by a 600 bp band in the *P. ulei* GCL012 DNA samples denominated T1, T2 and BMK1 (Figure 2-6, lines 2,3,4). The ITS sequence (Annex 1) aligned with an identity of 99.9% against *Microcyclus ulei* ITS sequences available in the NCBI database. Moreover, we amplified a band of 175bp from the T2 and BPK1 and DNA samples, which corresponded to the *P. ulei* specific partial region of TDF-Mu1 gene fragments (Figure 2-6, Lines 6 and 7). The microbiological and molecular results allowed us to confirm the purity and identity of the *P. ulei* GCL012 isolate.

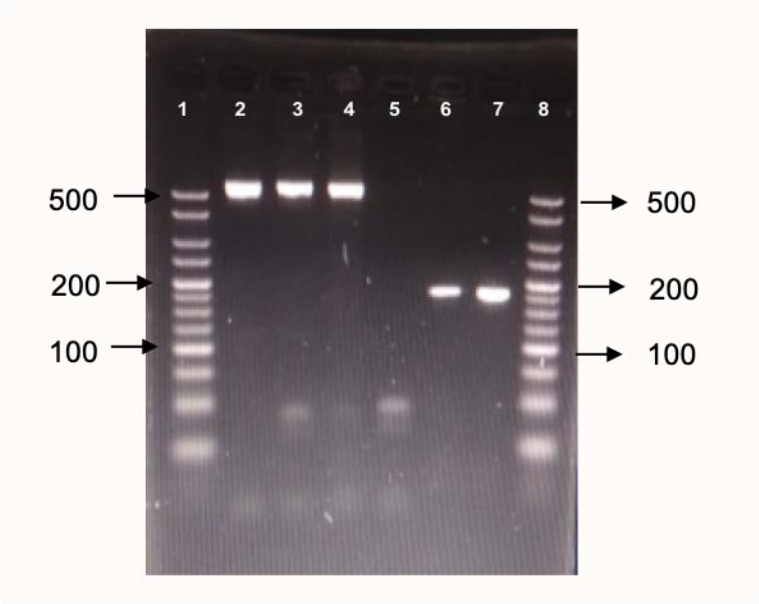


Figure 2-6. Electrophoresis in agarose gel of ITS and TDF-Mu genes amplimers: lines 1 and 8, hyper ladder 25bp marker Bioline®. Lines 2 - 4 corresponds to ITS amplimers from T1, T2 and BMK1 DNA samples and lines 5 - 7 correspond to TDF-Mu gene amplimers from T1, T2 and BMK1 DNA samples, respectively.

2.3.4. DNA sequencing yield

Illumina platform dataset was the first sequencing source generated in this project. We generated 18'894'474 short paired end (2x250 bp) illumina reads, which translates to a coverage of 100x for the theoretical *P. ulei* genome size (see below). With PacBio we generated a total of 2'019'638 raw reads, with 10'488'479'080 bp (for an estimated 111x coverage), with an N50 of 5'103 and a mean length of 5'193 bp. For the MinION, the flow cell produced 3'606'462 raw reads, with 7'152'783'610 bp, (for an estimated 76x coverage), with an N50 of 4'506 bp and a mean length of 1'983. The CCS content in all PacBio datasets was less than 1.7%, so we decided to use raw-reads for the genome assembly process.

2.3.5. Illumina reads pre-processing

In general, illumina raw reads showed good quality parameters in terms of adapter sequences representativity, GC content, per base sequence content, length distribution, duplication levels and overrepresented sequences. However, the reads presented low per base quality in both extremes, which is expected for this kind of technology (Figure 2-7). The FastQC results after quality trimming (leading-3, trailing-6) increased the reads quality, leading to a better distribution of the reads quality per position, which can be observed in the box whisker where the highest quality values are shown in green (Figure 2-7). A total of 392'593 low quality reads were eliminated by trimmomatic, reducing the number of reads from 18'894'474 from 18'501'881.

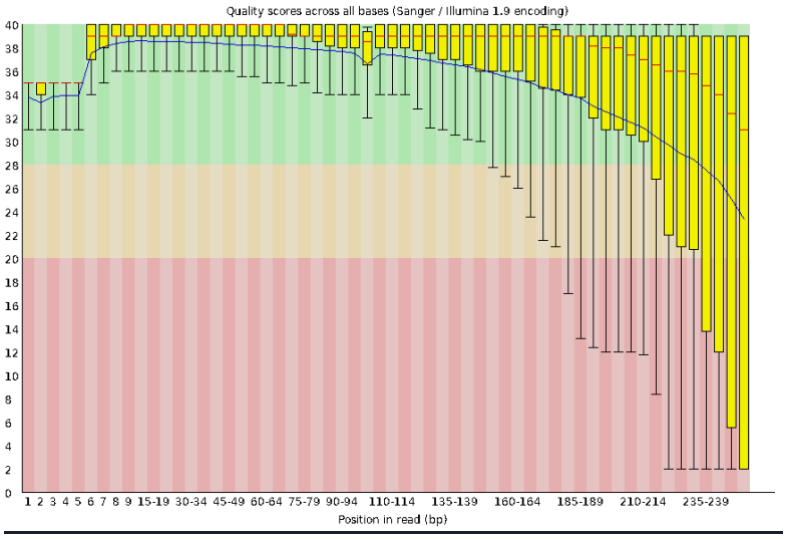


Figure 2-7. FastQC BoxWhisker graphs before the trimming

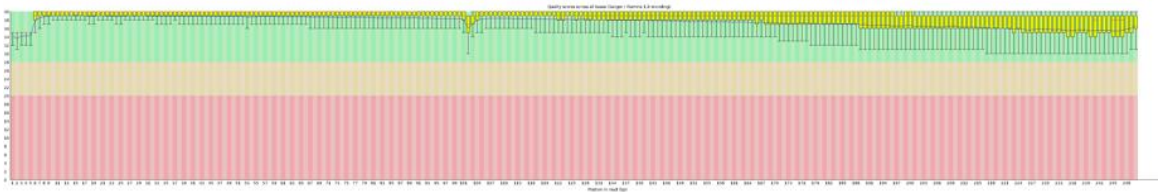


Figure 2-8. FastQC BoxWhisker graphs after trimming process

With respect to contaminant sequences, blobtools results showed the presence of 0.2 Mb of bacterial sequences in the illumina reads (Figure 2-8), identifying sequences from *Arthrobacter*, *Cupriavidus*, *Glutamicibacter*, *Herbaspirillum*, *Microbacterium*, *Nocardia* and *Rubrobacter* genus. Approximately 10% (2'009'653) of the total raw reads were eliminated in this phase, including *P. mori* mitochondrial genome mapped reads. The preprocessing of illumina reads lead a remnant of 16'884'821 short reads with high quality belonging to *P. ulei* which were the input to short reads assembler.

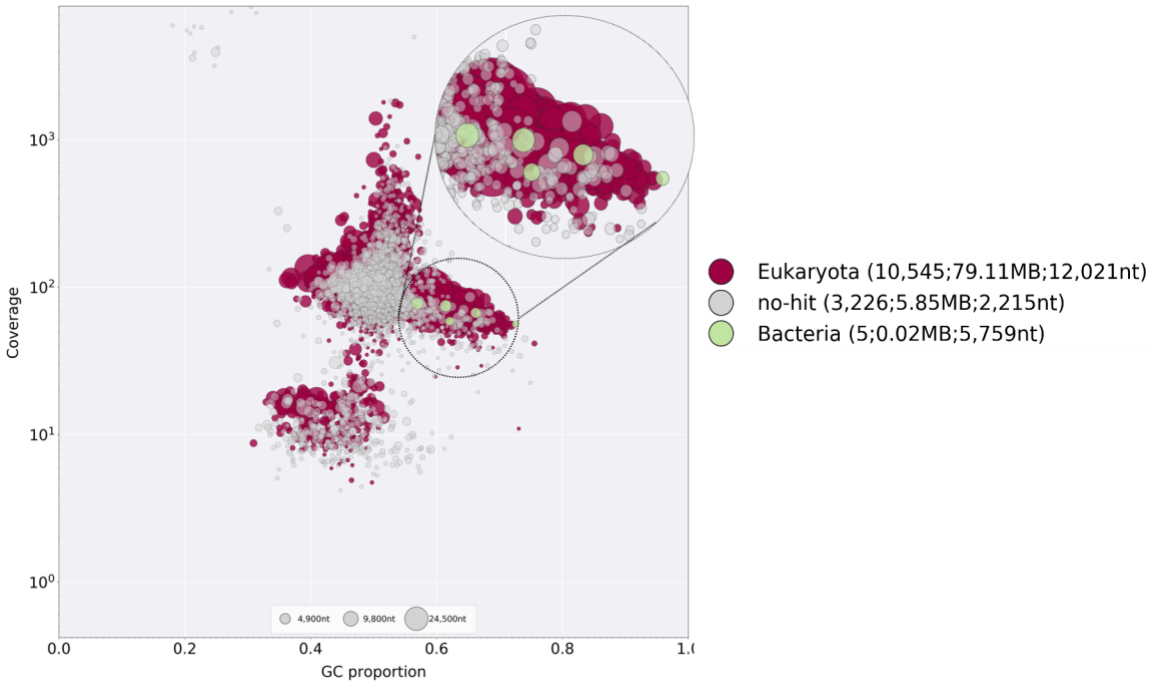


Figure 2-8. Bacterial sequences identified by BlobTools in the Illumina dataset

2.3.6. *kmer* spectrum of illumina sequencing reads

We exploited the illumina dataset to get an assembly-independent estimation of *P. ulei* genome characteristics, using the *kmer* spectrum calculated with *kmers* of length 21bp. With this, we estimated that the *P. ulei* genome size is around 75'152'984 bp, 36% of which were categorized into unique sequences and 63% as repeat content. This analysis also revealed that the data fitted well with a haploid genome (Figure 2-9). Based on these results, long reads sequencing platforms depth was calculated to cover 100x the estimated genome size.

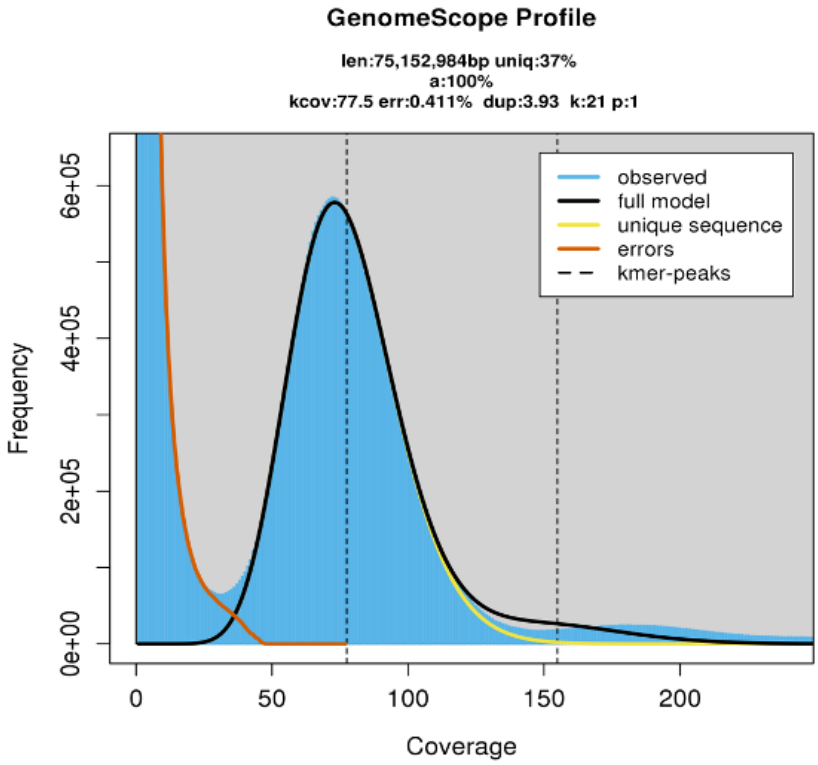


Figure 2-9. GnomeScope *K-mer* frequency distribution diagram

The *P. ulei kmer* spectrum results contributed with significant information useful to the sequence strategy design. *P. ulei* theoretical genome size, predicted by the GenomeScope program was very close to *P. fijiensis*, our reference species (75.1 Vs. 74 Mbps). Furthermore, the inference of a high repetitive content in *P. ulei* genome (47 Mbps) was determinant to increase the depth on long reads sequencing technologies which could better solve these hardly assembly regions.

2.3.7. Genome assembly

We carried out nine genome assemblies constituted by, i) two short reads assemblies, obtained with the SoapdeNovo2 and MaSuRCA assemblers. ii) two hybrid assemblies built with short and long reads using MaSuRCA (PacBio and Illumina data) and SPAdes (PacBio, Nanopore, and Illumina data) assemblers. ii) four long non-hybrid assemblies, three were built only with PacBio reads using

Wtdbg2, Falcon, and Canu assemblers, and one with ONT reads using Canu assembler. And iv) one long hybrid assembly built with both long reads datasets (PacBio and ONT) using Canu assembler. The contiguity and completeness metrics from every genome assembly strategy assessed in this project are presented in Table1- 4.

2.3.7.1. Short reads assemblies

Due to the large size and repeat content of *P. ulei*' s genome, assemblies exploiting only short reads could not generate suitable quality assemblies. Although Illumina sequences presented an excellent quality and yield, these assemblies showed, in general, a high level of fragmentation. SOAPdenovo2 generated the smallest genome assembly with 50.3 Mbps, in 36'541 contigs with an N50 value of 1'623 (Figure 2-10, Illu_Soap). We tried without success to improve the SOAPdenovo2 assembly using further *kmer* values, including 147mer, which Kmergenie predicted as the optimal *kmer* value for this dataset. MaSurCa offered better assembly results on the Illumina data alone, generating 84.9 Mbps assembly (Figure 2-10, Illu_Masurca) in 13'776 contigs and an N50 value of 11'182, almost seven times higher than with SOAPdenovo2 (Table 1-4). The genome completeness metrics revealed better results for Illu_Masurca assembly, identifying 35 complete BUSCOs that were misleading the Illu_Soap assembly (1'191 vs. 1'226 total BUSCOs identified).

The low performance of short reads assemblers in a genome reconstruction could be associated with the genome characteristics or sequencing technology (Li et al., 2011). The *kmer* spectrum distribution analysis of *P. ulei* identified repetitive sequences representing 63% of the *P. ulei* theoretical genome size. We concluded that the level of fragmentation of short reads assemblies are due to the high repetitive content in the *P. ulei* genome, which is described as the primary hurdle in the *de Bruijn graph* construction as overrepresented *kmers* increase the graph complexity and, subsequently, the consensus definition (Zimin et al., 2013). The SoapDeNovo2 complexity graph reduction phase, where the low coverage links and tiny repeats are removed, and the complex cycles generated by repetitive *kmers*

(bubbles) are solved, could lead to the genome consensus size underestimation (Luo et al., 2012), as was observed in the Illu_Soap genome, which was approximately 40 Mbps shorter than the genome sizes obtained integrating long reads to the assembly strategy.

On the other hand, the MaSuRCA assembler reduced the number of contigs in more than half compared with Illu_Soap assembly, increasing the N50 value and genome size. The better performance of MaSuRCA assembling *P. ulei* short reads could be attributed to the reduction of data complexity with the super-reads implementation in the genome reconstruction. The super-reads are originated through *kmer* information extending base by base as long as the extension is unique, this methodology reduces the lost information as each illumina read will be contained in a super-read (Zimin et al., 2013). These differences at algorithmic level could be associated with the improvement in the genome size generated by this strategy.

As we previously described, the low performance of short reads assemblers is attributed to *P. ulei* genome complexity. This could be confirmed with the short reads genome assembly results obtained for smaller and less repetitive genomes from *Pseudocercospora* genus. The *P. musae* (60 Mbps) and *P. eumusae* (47 Mbps) reference genomes were assembled with SOAPdeNovo using 22.7 and 26.0 million of illumina reads. However, these genomes presented better N50 values with 40.000 and 112.000 bp, respectively (Chang et al.,2016), exceeding it more than twice compared with Illu_Masurca assembly. With respect to the completeness metrics, the BUSCOs profiles were very similar between *P. ulei* (MaSurCa) and *P. musae* and *P. eumusa* genomes recovering, 93.3%, 94.2%, and 98.7% complete BUSCOs, respectively. From these last results, we inferred that gene occupied genome regions were efficiently reconstructed and that unsolved regions are associated with other genome features.

2.3.7.2. Hybrid assemblies

The incorporation of long reads to the assembly of the *P. ulei* genome greatly improved the contiguity and completeness metrics of the assemblies obtained with hybrid methods (Table 1-4). *P. ulei* hybrid genome assembly strategies generated with MaSuRCA (Illu_PacB_Masurca) and SPades software shown a dramatic increase to the contiguity achieving genome sizes of 85.6 and 92.6 Mbps with a lower number of contigs 1'193 and 3'804, respectively (Figure 2-10, Illu_PacB_Masurca and Spades). Illu_PacB_Masurca assembly increased the N50 value nine times compared to Illu_Masurca, which used only short reads.

The comparison between hybrid assemblers revealed that the genome contiguity was similar for the assemblies generated by both strategies, presenting N50 values of 110'987 and 147'130 bps, respectively. Nevertheless, Illu_PacB_Masurca genome size was 7 Mbps smaller than SPAdes assembly (Table 1-4). The MaSURCA assembler performance improved when long reads were integrated to the assembly strategy, however, Illu_PacB_Masurca small genome size could be associated with the long reads error rates which could interfere in the super-reads construction accuracy, and the overrepresentation of repetitive sequences in both datasets which leads to a merge of unsolved regions during the *de Bruijn graph* construction.

The Spades assembler presented better performance than all the assemblers that included short reads for the consensus's reconstruction. Even though the high number of contigs, Spades assembly showed high completeness metrics represented by 97.4% of complete BUSCOs with only 1.0% and 1.6% of fragmented and missing orthologs. The Spades genome assembly quality could be attributed to the *kmer* length adaptation's versatility during the de graph construction. The SPAdes assembler implements a modified *de Bruijn graph* data structure called multisized *de Bruijn graph,* which allows for varying the *kmer* length. Thus, this assembler implements smaller *kmers* in low coverage regions, reducing the fragmentation level, and larger kmers values in genomic regions with high coverage reducing the repeat sequences collapse (Bankevich et al., 2012).

2.3.7.3. Long reads assemblies

Assembly strategies only exploiting long reads generated genomes larger than 90 Mbps except for Falcon software (83.7 Mbps) (Table 1-4). The PacB_Wtdbg2 assembly was the most fragmented genome than all the assemblies generated from long reads (Table 1-4). The PacB_Wtdbg2 assembly was reconstructed into 1'311 contigs with an N50 value of 143'999 and a total size of 92.6 Mbs; these results positioned the Wtdbg2 assembler as the less efficient assembling *P. ulei* long reads. Wtdbg2 combines the flexibility of the OLC algorithm and the computational efficiency of the *de Bruijn graph*. This assembler implements a variant of the d*e Bruijn graph* termed fuzzy *de Bruijn graph* for genome consensus reconstruction. In this strategy, the reads are shopped into 256 pbs, and then they are aligned (all against all), reducing two overlapped reads into a bin. Therefore, the graph is constructed using bins as vertices and connecting them with an edge defined by the adjacency degree. (Ruan & Li, 2019). All the reconstructed assemblies using the traditional or a variation of *de Bruijn graph* data structure results allowed us to infer that the repetitive content in short and long sequenced reads is the main reason these assemblers cannot appropriately solve the *P. ulei genome* below this strategy.

Hierarchical assembly methods (with or without combining the two long sequencing datasets) presented the best performance assembling long reads datasets. From them, falcon software exhibited the lowest performance reconstructing *P. ulei* genome (Table 1-4). Even when the number of contigs was approximately half of the PacB_Wtdbg2 assembly, PacB_Falcon assembly was the smallest long reads assembly with 83.7 Mbs. Additionally, PacB_Falcon genome presented the lowest completeness metrics from the whole genome assembly strategies assessed in this project, assigning only 70.9% of complete BUSCOs, and the highest number of fragmented and missing BUSCOs with 158 and 225 genes, respectively (Table 1-4).

Our results demonstrated that hierarchical assembly methods used by Canu software improved the *P. ulei* genome contiguity in three orders of magnitude compared with short read assemblies (Table 1-4), We observed remarkable differences among the long non-hybrid assemblies reconstructed with canu using the nanopore or PacBio reads independently. Although Nano_canu assembly showed a smaller number of contigs than the PacB_canu assembly, its N50 value (1.4 Mbps) was half of the obtained PacB_canu exhibiting a small genome assembly size (Table 1-4). We hypothesized that the error rate of Nanopore reads could not be efficiently solved by the Canu algorithm's correction phases, since the ONT low accuracy (80%) affects the average identity across overlapped reads during the layout construction (Koren & Phillippy, 2015). We validate this hypothesis with the orthologous gene composition in Nano_Canu assembly, which presents a high level of fragmented and missing BUSCOs (Table 1-4).

The ONT technology can produce longer reads than PacBio, however, nanopore reads errors are more complex leading less complete and contiguous assemblies (Magi, Giusti & Tattini, 2016; Rang, Kloosterman & Ridder, 2018). Canu is more efficient assembling PacBio reads or can offer comparable results only when the ONT information is generated from both DNA strand. using D2 flow cell (Koren et al., 2017). Unfortunately, we only generated *P. ulei* nanopore reads with D1 flow cells obtaining information from one DNA strand. This reduced the canu software performance in both hybrid_long assembly (PacB_Nano_Canu) and also non_hybrid long assembly using ONT reads alone. Even though, Canu assembler was designed for the efficient assembly of single-molecule sequencing technologies, dealing with the high error rates presented in these technologies, it is still necessary technological developments in ONT sequencing and improved algorithms for their base calling and alignment (Koren & Phillippy, 2015; Chen et al., 2021).

Table 1-4. Quast and BUSCOs results for the non-hybrid, hybrid and hierarchical, genome assemblies strategies.

Assembly	Short reads assemblies		Hybrid assemblies		Long non-hybrid assemblies				Long hybrid
	(Only Illumina reads)		(Illumina, Pacbio and ONT reads)		(Only PacBio reads)			(Only ONT reads)	(PacBio and ONT reads)
Assembly name	Illu_SOAP	Illu_Masurca	Illu_PacB_Masurca	SPAdes	PacB_Wtdbg2	Hierarchical assemblies			
						PacB_Falcon	PacB_Canu	Nano_canu	Nano_Pab_canu
STATISTICS WITHOUT REFERENCE									
# Contigs	36.541	13.776	1.193	3.804	1.311	573	231	167	336
Longest Contig	54.712	97.624	911.021	1.095.545	617.313	2.072.718	11.256.605	4.202.079	5.800.913
Total length	50.382.058	84.979.727	85.638.836	92.612.415	92.666.995	83.710.889	93.812.866	91.182.491	95.849.438
N50	1.623	11.182	110.987	147.130	143.999	280.044	2.355.240	1.465.888	2.123.084
GC (%)	51.4	50.16	50.58	50,19	50.23	50.71	50.27	50,19	50,27

BUSCOs ASSIGNATIONS																		
Category	**Nº**	**%**	**Nº**	**%**	**Nº**	**%**	**Nº**	**%**	**Nº**	**%**	**Nº**	**%**	**Nº**	**%**	**Nº**	**%**	**Nº**	**%**
Complete BUSCOs	1.191	90,5	1.226	93,3	1.228	94,4	1.281	97.4	1.205	91,7	932	70,9	1.282	97,5	1.090	82,9	1.260	95,8
Complete single copy	1.188	90.3	1.225	93.2	1.208	91.9	1.280	97.3	1.204	91.6	932	70.9	1.273	96.8	1.089	82,8	1.244	94,6
Complete single duplicated	3	0.2	1	0.1	20	1.5	1	0.1	1	0.1	0	0.0	9	0.7	1	0,1	16	1,2
Fragmented	94	7.1	58	4.4	17	1.3	13	1.0	51	3.9	158	12.0	13	1.0	100	7,6	23	1,7
Missing	30	2.4	31	2.3	70	5.3	21	1.6	59	4.4	225	17.1	20	1,5	125	9,5	32	2,5
Total BUSCO groups searched	1315																	

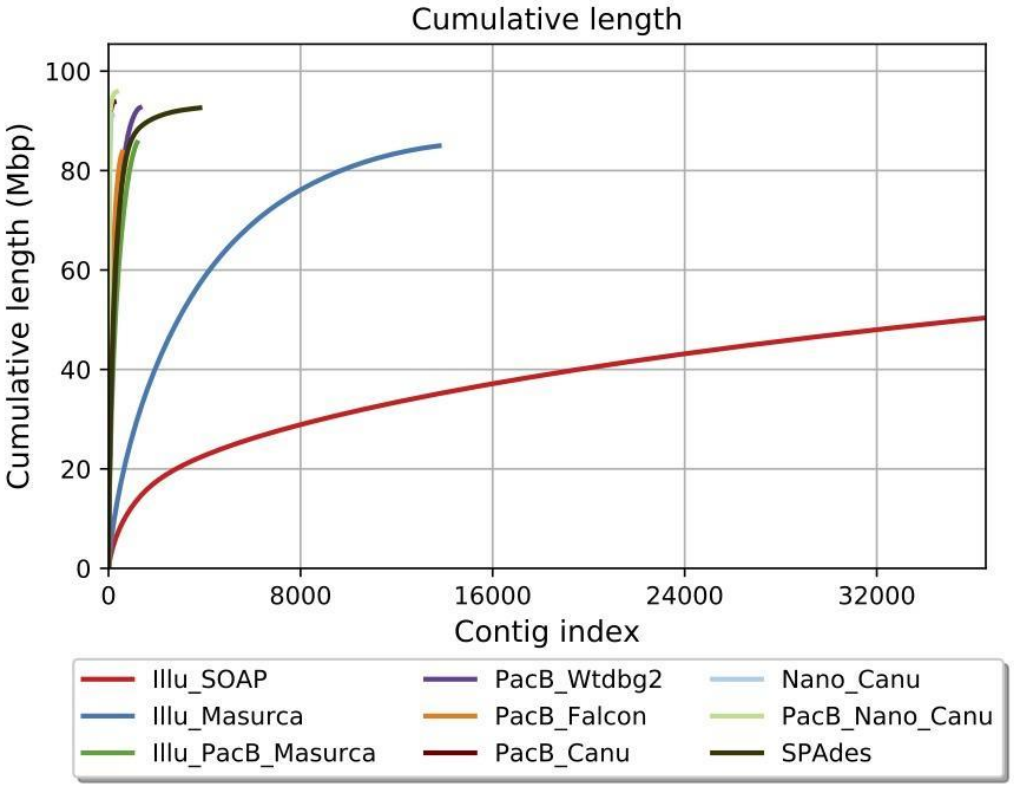


Figure 2-10. Cumulative length generated by Quast software for: short reads genome assemblies (Illu_SOAP, illu_Masurca), hybrids with short_long assemblies (Illu_PacB_Masurca, SpaDES), and long reads assemblies (PacB_falcon, PacB_Canu, PacB_SMRTLink, PacB_Nano_Canu and Nano_Canu.

The best assembly, in terms of contiguity and gene completeness, was generated by CANU v1.8 using the PacBio dataset. The assembled *P. ulei* genome has 93.8 Mbps and was fragmented into 231 contigs, where the largest was 11.2 Mbps, (N50 = 2.35 Mbps) and a GC% content of 50.27% (Table 1-4). BUSCO analysis identified 1'282 complete orthologous genes (97.4%) in the assembled genome. From these, 1'273 were single-copy, and only 0.7% were duplicated, supporting that this genome is haploid (Table 1-4). One of our more relevant findings is the *P. ulei* size genome was bigger than the theoretical size predicted from the *K-mer* frequency analysis with 93.8 Mbps, *P. ulei* genome is the largest genome so far reported in the Mycosphaerellaceae family (Figure 2-11) and in *Pseudocercospora spp* genus.

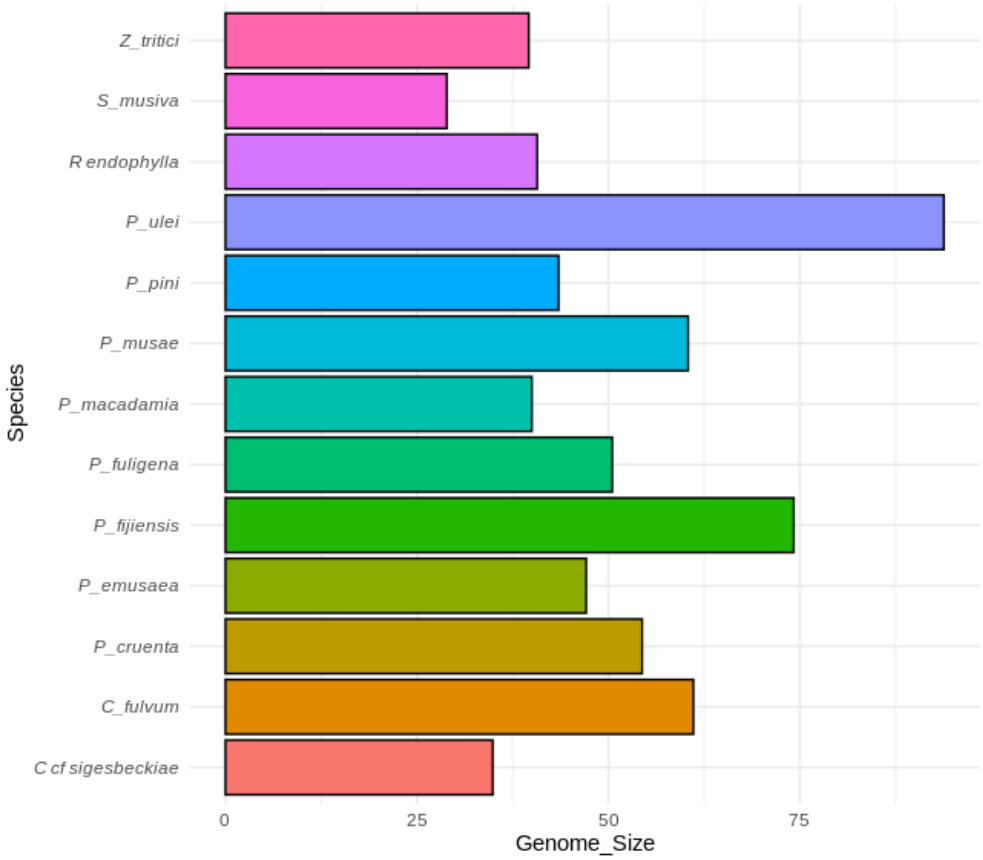


Figure 2-11. Comparison of *P. ulei* Genome size with other species of Mycosphaerellaceae family

Rusts fungi exhibits the biggest genome sizes of the fungi kingdom, this phenomenon have been correlated with their obligate trophy and with the ability to infect several hosts, which is conferred by the differential spore states during their life cycle (Jaswal et al., 2020). Similarly to the rust, *P. ulei* presents three types of spores in its life cycle, ie., conidia, spermagonia and ascospores (Hora et al., 2014), is plausible that the genome size expansion be related with its holimorphic nature.

2.3.8. Genome scaffolding

Pyscaf presented better performance with ONT raw reads than with the corrected data. The number of contigs revealed that 59 contigs were merged, increasing the N50 value from 2.35 to 2.82 Mpb and the largest contig length from 11.5 to 14.6 Mpb (Table 1-5). However, the genome size decreased 5Mpb with this scaffolding method compared with the Pacb_canu genome (Table 1-5), affecting the BUSCOs composition, which lost 84 complete single-copy genes previously identified in the non-scaffolded genome. This discrepancy could be associated with spurious alignment events led by the high error rate of ONT reads. We confirmed this, determining that only five junctions (contigs_1, contigs_2, contigs_3, contigs_30, and contigs_13) from the 59 were supported for more than 30 reads (Figure 2-12), this is a low number of reads considering that this sequencing technology covered 76x the *P. ulei* genome size.

The genome scaffolding strategy based on *kmer* paired generates genomes with better quality than the obtained with the PyScaf program (Table 1-5). In general, all LINKs evaluations merge less than 20 contigs without affecting the largest contig length value. From the six assessments, LINKs4 were selected as the final *P. ulei* genome because it merged the higher number of contigs without affecting neither the genome length nor the BUSCOs composition (Table 1-5); this genome was scaffolded using with 21mer, separated by 7’500 bp with a sliding window of 5 bp. The LINKs4 scaffolding strategy processed seventeen scaffolding events increasing the N50 value to 2.82 Mbps. The BUSCOs profile presented a slight variation compared with the non_scaffolded genome (Table 1-5), mainly represented by the loss of three duplicates and the gain of three missing.

Table 1-5 Comparison of scaffolding strategies quality metrics (the last column shows the non_scaffolded *P. ulei* genome quality metrics.

Scaffolder	**PyScaf 1**		**PyScaf 2**		**LINKS 1**		**LINKS 2**		**LINKS 3**		**LINKS 4**		**LINKS 5**		**LINKS 6**		**PacB_Canu**	
Parameters	ONT corrected reads		ONT raw-data		Default Parameters		d 5000 k21		d 6000 t5 k21		d 7500 t5 k21		d 7500 t5 k35		d 7500 t5 k53		Choiced assembler	
STATISTICS																		
# Contigs	212		172		216		220		216		214		228		218		231	
Largest Contig	14.671.800		14.664.796		11.240.615		11.240.615		11.240.615		11.240.615		11.240.615		11.240.615		11.256.605	
Total length	90.658.122		88.105.399		93.709.547		93.706.565		90.658.122		93.730.151		90.658.122		93.715.734		93.812.866	
N50	2.352.044		2.827.303		2.655.028		2.655.028		2655028		2.827.303		2352044		2655028		2.355.240	
GC (%)	50		50		50		50		50		50		50		50		50.27	
BUSCOs																		
	Nº	**%**	**Nº**	**%**	**Nº**	**%**	**Nº**	**%**	**Nº**	**%**	**Nº**	**%**	**Nº**	**%**	**Nº**	**%**	**Nº**	**%**
Complete BUSCOs	1.250	95	1.195	90.9	1.281	97.5	1.281	97.5	1.280	97	1.280	97	1.280	97	1.281	97.5	1.282	98
Complete single copy	1.244	95	1.189	90.4	1.275	97	1.275	97	1.274	97	1.274	97	1.274	97	1.275	97	1.273	96.8
Complete single duplicated	6	1	6	0.5	6	0.5	6	0.5	6	1	6	1	6	1	6	0.5	9	0.7
Fragmented	16	1	12	0.9	11	0.8	11	0.8	12	1	12	1	12	1	11	0.8	13	1.0
Missing	49	4	108	8.2	23	1.7	23	1.7	23	2	23	2	23	2	23	1.7	20	2
Total number of BUSCOs	1.315																	

-d distance between *k-mer* pairs
-t step of sliding window of *k-mer* extraction
-k *k-mer* value (default -k 15)

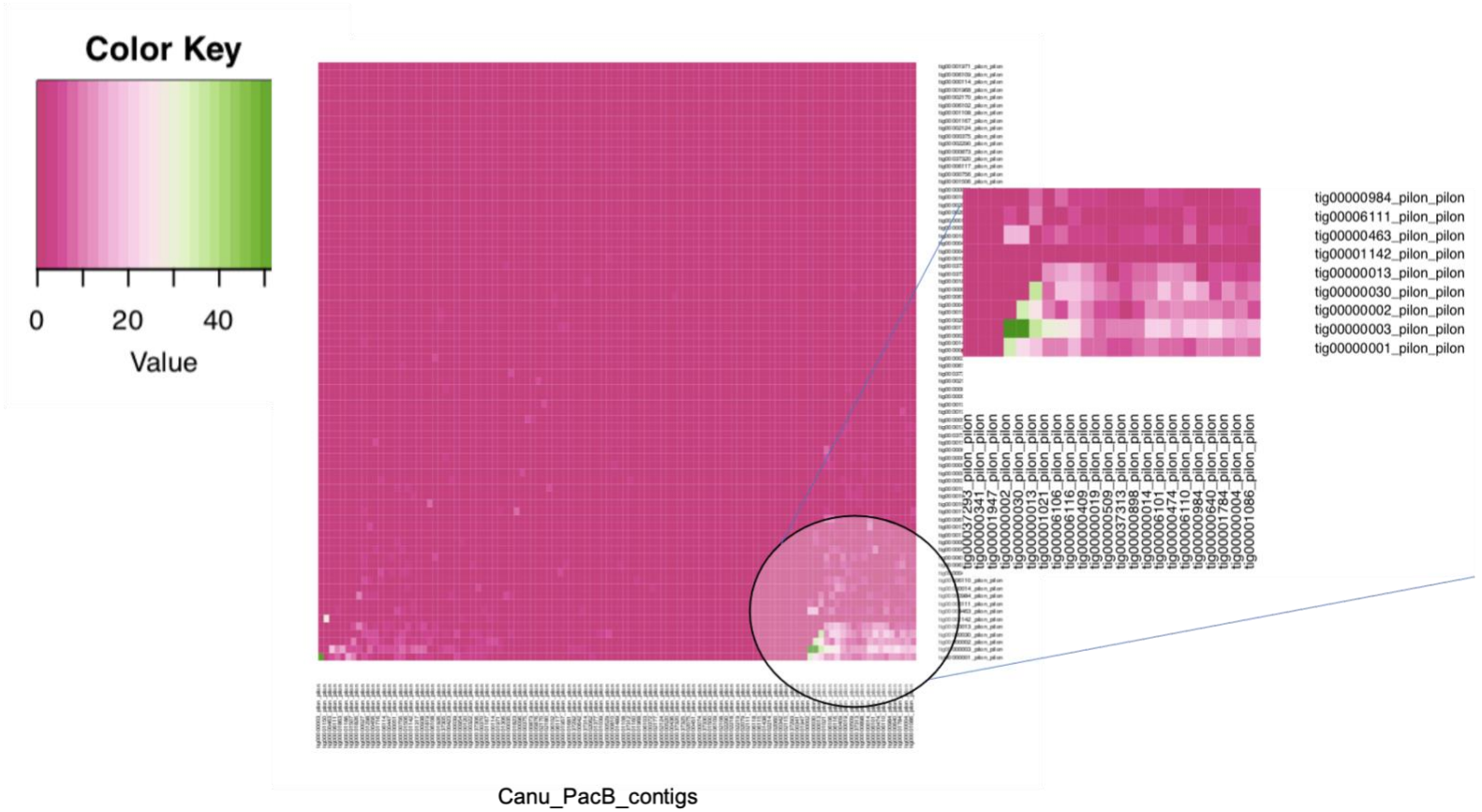
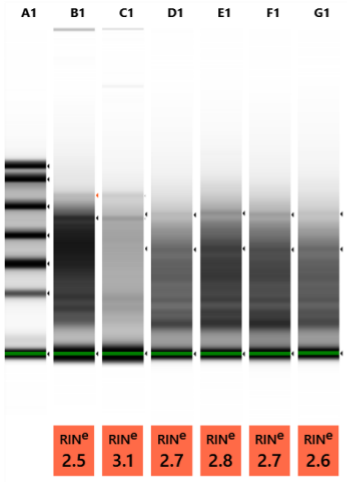


Figure 2-12. Heatmap, number of nanopore reads that merge two contigs.

2.3.9. Structural gene Annotation

2.3.9.1. *P. ulei* RNA extraction and sequencing

The *P. ulei* RNA extraction represent one of the biggest challenges in this project. Although we implemented low detrimental RNA extractions and RNA stabilization tubes to avoid the degradation during the sample shipments, we could not generate *P. ulei* RNA samples with acceptable RNA Integrity Number (RIN>=6) parameters. Despite the above, we decided to sequence the samples with better RIN values (E1=2.8 and C1=3.1) for the RNA sequencing reads production (Figure 2-13).



Shipping number	Sample ID	RNA Concentration
1	A1	84.9
	B1	66.9
	C1	67.0
	D1	77.3
	E1	106
	F1	90.1
	G1	71.3
2	B1	88.9
	C1	27.5

Figure 2-13. *P. ulei* RNA extraction integrity and concentration parameters

A total of 176’028’243 and 165’602’'186 strand-specific raw RNA reads with a median length of 149 bp were produced from both *P. ulei* early (C1=Pse_RNA_1) and later sporulation (E1=Pse_RNA_2) stages, respectively (Table 1-6). FastQC results revealed that the RNAseq data did not accomplish the minimal quality values for adapter content, GC content per sequence, and sequence duplication levels. The adapter sequences were overrepresented in *P. ulei* RNAseq data. The adaptors elimination process reduced the number of raw reads to 56% and 60% for Pse_RNA_1 and Pse_RNA_2 RNA dataset, respectively. These results could be

associated with *P. ulei* RNA's low integrity which could have affected the RNA library preparation. Conversely, the ribosomal sequences' representativity was less than 5% in both datasets (Table 1-7). The raw data pre-processing resulted in 90’051’015 and 103’057’749,00 reads which were used for downstream processes (Table 1-6).

Table 1-6 Number of RNAseq reads in different processing stages.

Library	Number of raw_reads	After adaptors elimination	After of ribosomal elimination
PseuRNA1_R1.1P	165’602’186	94’123’974	90’051’015
PseuRNA1_R2.1P	165’602’186	94’123’974	90’051’015
PseuRNA2_R1.1P	176’028’243	106’698’485	103’057’749
PseuRNA2_R2.1P	176’028’243	106’698’485	103’057’749

The preprocessed *P. ulei* RNA reads presented good mapping rates when they were aligned to the *P. ulei* genome (Table 1-7). More than 90% of the two datasets mapped concordantly at least once to the reference, and less than 6% did not align to the *P. ulei* genome (Table 1-7). These results show that despite the loss of more than 50% of RNAseq raw reads during the adaptors removal, the remnant data could represent an optimal evidence resource to *P. ulei* structural annotation enrichment.

Table 1-7: Mapping stats of *P. ulei* RNA preprocessed reads against *P. ulei* genome

HISAT2	Pse_ulei_1	Pse_ulei_2
Number of reads	90’051’015	103’057’749
Mapped concordantly 0 times	3.20	2.82
Mapped concordantly 1 times	91.28	94.63
concordantly concordantly >1	5.52	2.55

The introns evidence was also enriched with *P. fijiensis* RNAseq information available in the NCBI database. *P. fijiensis* transcripts data was constituted by three

single-end RNA datasets (Pse_fijen_1, Pse_fijen_2, and Pse_fijen_3) with content 28’280’995, 37’521’'028, and 23’316’046 raw reads, respectively. The alignments of *P. fijiensis* RNA seq data against the *P. ulei* genome presented low mapping rates. Only 1.8, 2.3, and 3.28% reads from the three *P. fijiensis* RNAseq datasets were aligned concordantly at least once, while 16.6, 19.2, and 30.9% were mapped concordantly more than once (Table 1-8). Despite the unmapped reads rates, which were higher than 69% for all the *P. fijiensis* RNAseq datasets (Table 1-8), we decided to include *P. fijiensis* transcripts evidence in the *P. ulei* genes prediction process.

Table 1-8: Mapping stats of *P. fijiensis* RNA preprocessed reads against *P. ulei* genome

HISAT2	*Pse_fijiensis*_1		*Pse_fijiensis_2*		*Pse_fijiensis_3*	
	N	%	N	%	N	%
Number total of reads	28280995		37521028		23316046	
Mapped concordantly 0 times	23577755	83.37%	30294276	80.74%	16096313	69.04%
Mapped concordantly 1 times	532859	1.88%	870981	2.32%	764582	3.28%
Concordantly concordantly >1	4170381	14.75%	6355771	16.94%	6455151	27.69%
Overall alignment rate	16.63%		19.26%		30.96%	

2.3.9.2. Gene Prediction Optimization

In general, the structural genes annotation presented better performance when training evidence from several sources i.e., protein sequences, transcripts sequences (RNAseq), and orthologous gene coordinates from the genome assembly were provided to the annotation pipeline (Figure 2-14). In the next paragraphs we will describe the contributions of the named evidence resources (obtained from sister species as well as, from *P. ulei*) in the gene prediction.

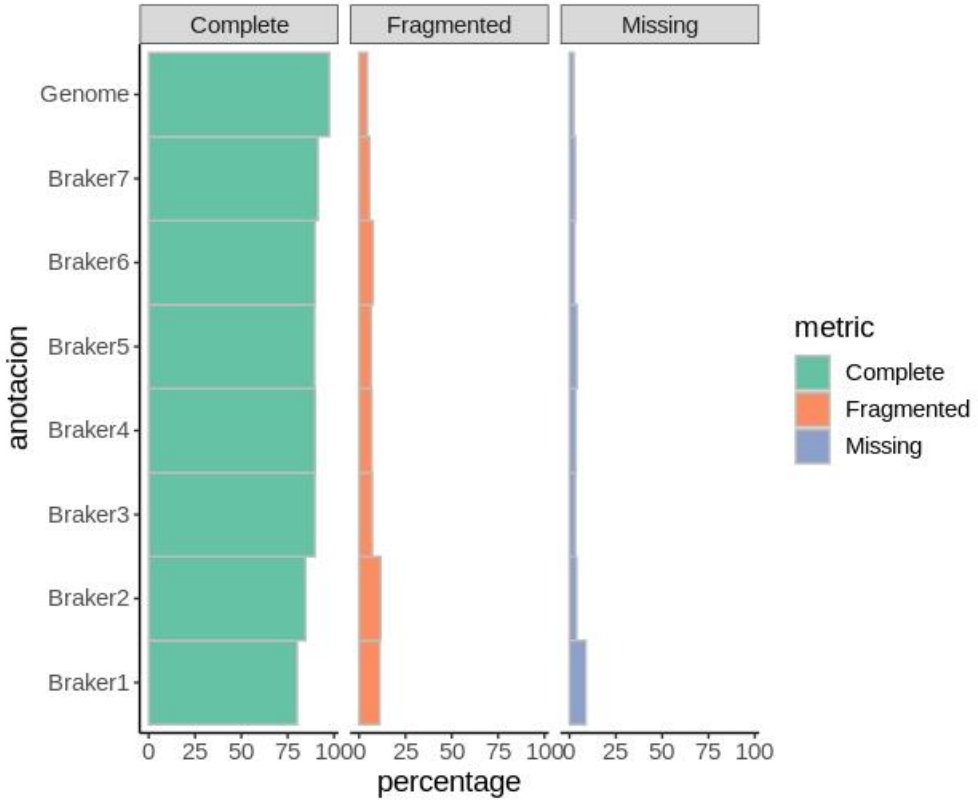


Figure 2-14. *P. ulei* RNA extraction integrity and concentration parameters

2.3.9.2.1. Contribution of protein sequences evidence to the structural annotation

The gene prediction process using only *Pseudocercospora fijiensis* protein catalog represented by Braker1 identified 10’090 genes containing 23’227 exons and 13’137 introns. Although, the number of exons and introns predicted in Braker1 annotation were the lowest from all annotation pipelines, the predicted genes presented a complete structure. It was confirmed with the GAG program that identified 10’078 complete CDS, from which only twelve were incomplete, identifying six CDS without start codon and six without stop codon (Table 1-9).

Conversely, the integration of *Pseudocercospora* spp protein sequences in Braker2 had a negative effect on the annotation performance. Even though this annotation pipeline predicted the highest number of genes of all the annotations with 18’839 genes (Table 1-9), the predicted genes presented a dramatic reduction of complete

CDS predicted by GAG program, identifying that all the 11’263 CDS identified did not have start and either stop codon.

The predicted genes from both annotation pipelines that implement protein sequences as evidence exhibited the lowest gene completeness metrics from all the performed annotations pipelines. Braker1 and Braker2 BUSCOs profiles were represented by 80.1 and 84.3% of complete BUSCOs, respectively, and the number of fragmented and missing BUSCOs was overrepresented in both annotations presenting 11.1% and 11.6% of fragmented BUSCOs and 8.8 and 3.8% of missing BUSCOs, respectively (Table 1-9).

Due to amino acid sequence conservation level among species, gene predictors can be efficiently trained with protein sequences from taxonomically related species, providing evidence for the prediction of exons (Yandell & Ence. 2012; Hoff & Stanke. 2015). However, the species from which the protein evidence to be used for the gene predictors training should be selected carefully, due to the low similarity rates between proteins could affect the predictors' accuracy (Koonin & Galperin, 2003; Keller et al.,2008).

Considering the species richness in the *Pseudocercospora* genus, we hypothesized that the so far *Pseudocercospora spp* sequenced species, could not be the *P. ulei* closest sister species. It is plausible that differences at the protein sequence level given by the taxonomic relationship distant between these three species could affect the predictor's accuracy in Braker2 annotation, generating a biased exons prediction. It was also confirmed with the incorporation of the proteomes from two Mycosphaerellaceae family fungal models (*Cladosporium fulvum* and *Zymoseptoria tritici)* to the protein sequences evidence in Braker5 pipeline, which reduced the number of predicted genes decreasing the annotation completeness metrics (Table 1-10).

Finally, it has been described that protein sequence does not provide evidence about splicing sites and intron sequences, leading to incomplete annotations (Hass et al., 2011). The low completeness metrics of Braker1 and Braker2 annotations could also be associated with the poor evidence of exon boundaries of the *Pseudocercospora* spp. protein sequences, which could have affected the optimal identification of complete genes during prediction.

2.3.9.2.2. Contribution of transcripts sequences and orthologous gene coordinates evidence to the structural annotation

A total of 10’858 genes composed of 34’259 exons and 23’415 introns were predicted in Braker3 annotation after integrating *P. fijiensis* RNAseq to the annotation pipeline (Table 1-9, Braker3). The *P. fijiensis* RNAseq evidence, improved the predicted gene completeness, identifying 126 complete BUSCOs more than the *P. fijiensis* protein alignments guided prediction (Table 1-10). Additionally, Braker3 BUSCOs profile presented a reduction of 4% and 5.6% of the number of fragmented and missing BUSCOs, respectively, compared with Braker1 annotation (Table 1-9).

The RNAseq information is an evidence resource key in the gene prediction, it enables to increase the gene predictors accuracy through the identification of exon-intron junctions, as well as splicing, initiation, and termination transcription sites (Keilwagen et al,. 2018). Despite the improvement in the predicted gene completeness metrics in Braker3, the majority of the CDSs detected by GAG program does not have start and stop codons (Table 1-9). This result could be attributed to the low quantity of *P. fijiensis* RNAseq reads supporting *P. ulei* gene models (Table 1-8), leading to a poor exon-intron junctions evidence which could decrease the predictors accuracy. Therefore, we concluded that *P. fijiensis* transcripts hints were not enough for improving the prediction of complete CDSs (Table 1- 9).

Inversely, the implementation of *P. ulei* genome orthologous gene coordinates, increased the predictors performance improving both the number predicted genes and the annotation completeness metrics (Table 1-9, Braker4). A total of 11’089 genes composed by 31’867 exons and 20’790 introns respectively, were predicted in Braker4 annotation. The integration of *P. ulei* genome informed gene coordinates improved the predictors accuracy, significatively increasing the number of complete CDSs identified by GAG statistics (Table 1-9). In detail, GAG identified 11’068 complete CDSs, and nine incomplete, seven CDS without start codon and two without stop codon. Regarding the predicted gene completeness, *P. ulei* orthologous coordinates slightly improved the BUSCOs profile, identifying three single copy BUSCOs more than Braker3 annotation, and recovering three fragmented and two missing BUSCOs more (Table 1-10). These results highlighted the importance of the implementation of authentic evidence from the annotated genome in the genes prediction, which improved the structural annotation increasing the predictors accuracy significatively (Table 1-9).

With respect to the contributions of *P. ulei* transcripts as evidence for the annotation pipeline, *P. ulei* RNAseq allowed the prediction of 10’794 genes optimizing the exons and introns identification. GAG identified 9’868 exons and 8’967 introns more than Braker5 annotation (Table 1-9), conserving the trend in the number of complete and uncomplete CDSs. The Braker6 predicted gene completeness metrics were represented by 89.7% and 85.1% of complete and single copy BUSCOs (Table 1-10), respectively. One of the most relevant results of Braker6 annotation is the number of BUSCOs duplicated in the orthologous assignments. *P. ulei* transcripts evidence allowed for the identification of a total of 61 duplicated genes (Table 1-10). These results confirmed the importance of integrating authentic transcripts from the annotated genome (RNAseq *P. ulei*) as evidence for the structural annotation. The *P. ulei* RNAseq evidence increased the predictors sensitivity, optimizing the intron-exon boundaries identification, this result could be also associated with a higher coverage of splicing and introns sites for *P. ulei* RNAseq reads.

2.3.9.3. The effect of the genome masking in the structure annotation.

The structural annotation process often requires the genome masking to reduce the misprediction of CDS caused by repetitive sequences (Yandell & Ence, 2012). It has been well described that masked repetitive sequences optimize gene prediction (Ejigu & Jung, 2020). However, considering that the current project is addressed to identify *P. ulei* pathogenicity-related genes, which use to be localized in genomic compartments with high repetitive content (Dong, Raffaele & Kamoun, 2015), we decided to evaluate the effect of *P. ulei* unmasked genome in gene prediction performance given the whole evidence resources assessed in the *P. ulei* structural gene annotation.

The Braker7 annotation predicted the highest number of genes compared with the previous annotation pipelines, recovering 12'745 gene models, and maximizing the number of exons and introns to 59'250 and 46'513, respectively. From the whole predicted genes, GAG identified 12'700 complete CDSs and 37 incomplete, comprised by 15 CDSs without start codon, and 22 CDSs without stop coding (Table 1-9). Likewise, the Braker7 predicted genes completeness metrics were the best compared with all the previous annotation pipelines, presenting 91.3% of complete BUSCOs and the lowest percentages of fragmented and missing BUSCOs (Table 1-10). Because of the Braker7 annotation quality metrics in terms of completeness and number of predicted genes, we decided to use this annotation for the downstream analyses.

Table 1-9. GAG metrics for every annotation pipeline.

GAG metrics							
Category	**Braker 1**	**Braker 2**	**Braker 3**	**Braker 4**	**Braker 5**	**Braker 6**	**Braker 7**
# of genes predicted	10090	18839	10858	11089	9892	10794	12745
Number of mRNAs	10090	11263	10844	11077	9885	10786	12737
Number of exons	23227	35699	34259	31867	32054	41922	59250
Number of introns	13137	24436	23415	20790	22169	31136	46513
CDS complete	10078	0	0	11068	9873	10772	12700

CDS start, no stop	6	0	0	7	8	8	15
CDS stop, no start	6	0	0	2	4	6	22
CDS no start, no stop	0	11263	10844	0	0	0	0
% of genome covered by genes	16,2	21,1	20,3	19,9	20,6	21,4	38.6

Table 1-10. Predicted genes BUSCOs profiles

Annotation	Evidence	BUSCOs									
		complete		complete single		duplicated		Fragmented		missing	
		#	%	#	%	#	%	#	%	#	%
Braker 1	*P. fijiensis* proteins catalog	1054	80,1	1047	79,6	7	0,5	146	11,1	115	8,8
Braker 2	NRDB of *Pseudocercospora* proteins	1113	84,6	1085	82,5	28	2,1	152	11,6	50	3,8
Braker 3	NRDB_Pseudocercospora + **RNAseq *P. fijiensis***	1180	89,7	1171	89	9	0,7	93	7,1	42	3,2
Braker 4	NRDB_ Pseudocercospora + RNAseq *P. fijiensis* + **Augustus Busco coordinates from the scaffolded genome**	1181	89,8	1174	89,3	7	0,5	90	6,8	44	3,4
Braker 5	NRDB_ *Pseudocercospora* + RNAseq *P. fijiensis* + Augustus Busco coordinates from the scaffolded genome + **NRDB *Z. tritici* and *C. fulvum***	1177	89,5	1171	89	6	0,5	86	6,5	52	4
Braker 6	NRDB_ *Pseudocercospora* + RNAseq *P. fijiensis* + Augustus Busco coordinates from the scaffolded genome + ***P. ulei* RNAseq**	1180	89,7	1119	85,1	61	4,6	99	7,5	36	2,8
Braker 7	NRDB_ *Pseudocercospora* + RNAseq *P. fijiensis* + Augustus Busco coordinates from the scaffolded genome + *P. ulei* RNAseq	1201	91.3	1193	90.7	8	0.6	74	5.6	40	3.1
Genome	*P. ulei* scaffolded genome completeness metrics	1.280	97	1.274	97	6	1	12	1	23	2

* Added evidence source in bold.

2.3.9.4. Selection of the best *P. ulei* gene models

Although several evidence sources to optimize *P. ulei*'s structural annotation were provided, the predicted genes' completeness (BUSCOs profile) never reached the obtained for *P. ulei* genome assembly (Table 1-10). Thus, we decided to explore the consensus potential of the gene models obtained from the most complete structural annotations, as well as the orthologous genes present in the scaffolded genome, and the *de novo* transcripts aligned to the *P. ulei* genome as evidence resources (See methods) to choose a single prediction that better represents each putative gene.

The unique (complete and duplicated) BUSCOs identified from all the annotations allowed us to choose the Braker_6 predicted genes as the most complete. It recovered 16 undiscovered BUSCOs in other annotations, sharing 13 of them with the *P. ule*i genome BUSCOs profile (Figure 2-15, black arrows). The *P. ulei* genome orthologous assignments recovered several Orthologous non detected in the gene annotation (Figure 2-15, red circles). Therefore, we selected the predicted gene models from Braker6 and the orthologous genes annotated from the *P. ulei* genome for the best gene models selection.

With respect evidence of transcripts' alignments we first assessed the quality of both *de novo* and by reference assembly transcripts. The *de novo* assembled transcripts were represented by a total of 62.9359 transcripts, containing 95.8% of complete orthologous genes, 12.7 % of single-copy and 2.3 and 1.6% of fragmented and missing BUSCOs transcripts, respectively. Besides, *de novo* assembly strategy recovered a higher percentage of duplicated BUSCOs transcripts (83.1%) compared with the by reference transcriptome assembly (70.5%) (Table 1-11). The duplicated orthologous content could be associated with *de novo* transcriptome assembly's better performance in the gene isoforms identification. Contrastingly, the guided by reference transcriptome assembly exhibited exceeded two times the missing BUSCOs content and three times the fragmented. We decided to select the *de novo P. ulei* transcriptome assembly as an independent track for the gene models selection based on the described results.

Table 1-11. Quality metrics of transcriptome assembly strategies, (number of predicted transcripts and BUSCO profiles)

TRANSCRIPTOME ASSEMBLY	TRINITY	STRING-TIE
# of assembled transcripts	62'359	17'083
BUSCOs Complete	95.8%	89.4%
BUSCOs Single Copy	12.7%	18.9%
BUSCOs Duplicated	83.1%	70.5%
BUSCOs Fragmented	2.6%	4.2%
BUSCOs Missing	1.6%	6.4%

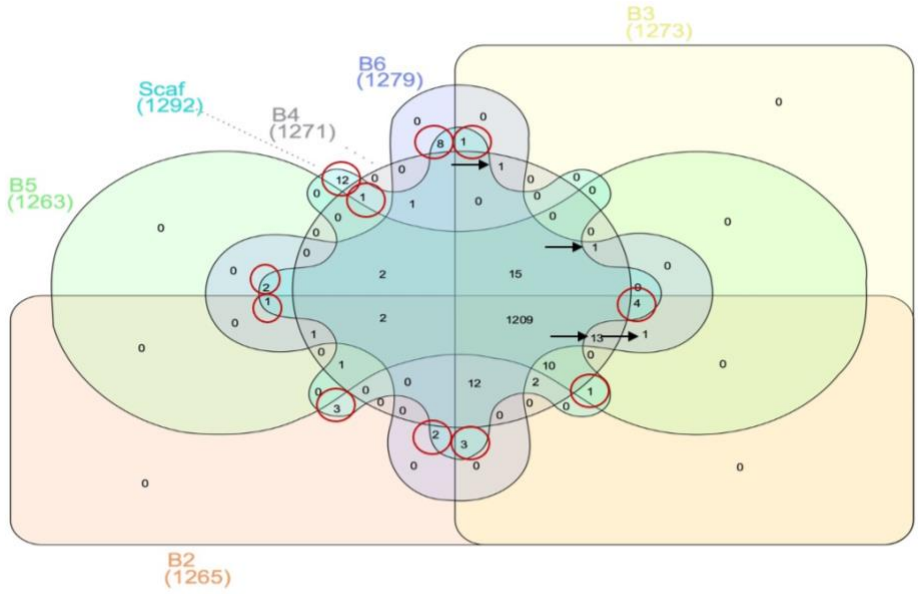


Figure 2-15. BUSCOs intersection among the six structural annotation pipelines and *P. ulei* scaffolded genome.

Despite the efforts to collect the gene models evidence, the EVM combiner could not efficiently select the best gene models, due to the non-canonical or cryptic splicing sites' overrepresentation in the *P. ulei* gene models assigned. The gene-splicing is defined by

the introns' remotion from pre- mRNA, generating an mRNA mature, subsequently translated into protein. Introns' excision is given by the precise recognition of a couple of conserved nucleotides in the 5' and 3' introns extreme. These sites are known as canonical splicing sites and are represented by GT combinations for the 5' end and AG for the 3' end (Frey & Pucker, 2020). Variations at the sequence level of these couple of nucleotides are known as the non-canonical splicing sites, which are frequently found in fungal genomes (Kupfer et al., 2004). Non-canonical splicing sites have an essential role in gene expression regulation and are the source for the emergence of new transcripts, the diversification of these sites is related to the transposable elements, especially those of the Alu class (Sibley, Blazquez & Ule, 2016).

Even though the identification of non-canonical splice sites could be associated with erroneous annotations generated by the lack of experimental gene validations and the presence of species-specific or non-conserved genes (Zhao et al., 2013). These results establish new research questions about whether the overrepresentation of these sites is given by incorrect annotations or if they are truly the source of *P. ulei* proteome diversity expanding.

2.3.10. Orthology analysis

The cluster analysis of the predicted proteins from the eight *Pseudocercospora* species revealed that 110'279 genes (87.4% of the total) were assigned to 4'132 orthogroups. The *P. ulei* clusters suggested that 48.5% (n = 6'184) of the proteins had orthologous conserved in the *Pseudocercospora* genus. Among the cluster of homologous genes, 3452 single copy ortholog sequences were detected, and *P. ulei* exhibited the highest number of species-specific genes with 77 sequences. In addition, *P. ulei* presented a higher number of unassigned genes to orthogroups with 6'561 genes corresponding to 58.5%. The orthology relationships among *the Pseudocercospora* genus suggest that the first seven *Pseudocercospora* species are all relatively closely related as they presented between 83.5 and 97.8% of genes assigned to orthogroups (Table 1-12). In contrast, *P.*

ulei is more distantly related to the sequenced fungus in this genus, presenting less than 50% genes assigned.

The orthologous gene clusters prediction is a source to assess the quality of the genome assembly and structural gene annotation (Emms & Kelly, 2015). In this context, conserved gene families' identification within the related species will be maximized with the genome completeness and the gene prediction accuracy (Hoff & Stanke, 2015). Nevertheless, we conclude that the cause for the low number of genes in orthogroups obtained for *P. ulei* could be a poor species sampling and not low quality in the assembly or the gene annotation. Probably, the *Pseudocercospora* species used in our comparative analysis were not the closest related species to *P. ulei.*

Table 1-12. Orthologous analysis of *P. ulei* with the sequenced *Pseudocercospora species*.

Species	Number of genes	Number of genes in orthogroups	Nº unassigned genes (%)	Nº of orthogroups containing species	Nº of species-specific orthogroups	Nº of genes in species-specific orthogroups (%)
P. eumusea	12632	11507 (91.1)	1125 (8.9)	10077 (89.6)	3	7 (0.1)
P. fijiensis	13066	10917 (83.6)	2149 (16.4)	10365 (90.0)	1	5 (0.0)
P. musea	13129	11570 (88.1)	1559 (11.9)	9673 (85.9)	7	23 (0.2)
P. cruenta	14228	13917 (97.8)	311 (2.2)	12173 (85.5)	2	7 (0.0)
P. fuligena	14494	14110 (97.4)	384 (2.6)	12137 (85.3)	3	8 (0.1)
P. macadamia	15430	13890 (90.0)	1540 (10.0)	11804 (82.9)	8	44 (0.3)
P. pini	14555	14237 (97.8)	318 (2.2)	12372 (86.9)	0	0 (0.0)
P. ulei	12745	6184 (48.5)	6561 (58.5)	5810 (40.8)	23	77 (0.6)

2.3.11. *P. ulei* phylogenetic replacement

The phylogenetic analysis based on the four genomic loci showed many deeply branching points without or with low support (Appendix 3), which could suggest missing taxa due to lack of diversity exploration, or taxonomic incongruence. The recent discovery of new *Pseudocercospora* species in sampling expeditions from tropical and subtropical environments (Bakhshi et al ., 2014; Silva et al., 2016) and the taxonomy studies (Quaedvlieg et al., 2012; Crous et al., 2013; Bakhshi et al., 2014; Nakashima et al., 2016; Silva et al., 2016), had revealed that the *Pseudocercospora* genus is highly diverse and

has not been fully explored. Nevertheless, we could identify four deeply branching clades with good support, Clades A to D (Appendix 3). Clade D is very well supported (Sh - aLRT=92.9/ UFBoots=97), and most of the species that have been explored genetically are located here (six genomes out of eight form *Pseudocercospora*, 160 species form this clade). *P. ulei*, could not be assigned to either of these four clades, but the topology suggests it is related to Clades B and C. The closest relative of *P. ulei* is *P. camelliicola*, a pathogen of camellia cultivars. This topology shows that *P. ulei* is among a group of *Pseudocercospora* that has not been genetically explored yet, and our *P. ulei* genomes is the first in this group (Appendix 3).

The orthologous gene phylogenomic analysis confirmed the previous findings, with *P. fijiensis*, *P. eumusae*, *P. musae*, *P. cruenta* and *P. pini_densiflora*, forming a clear clade (Clade D in the 4-marker phylogeny), and *P. ulei* and *P. macadamia* forming another one, and being sisters to each other (figure 2-16). These results strengthen the hypothesis that *P .ulei* is related to Clades B and C in the 4-marker phylogeny (Appendix 3). The genome of *P. ulei* contributes new genome information in a group of *Pseudocercospora* that is still not fully explored from the genomics point of view, and could help answer evolutionary and biological unsolved questions for these plant pathogens.

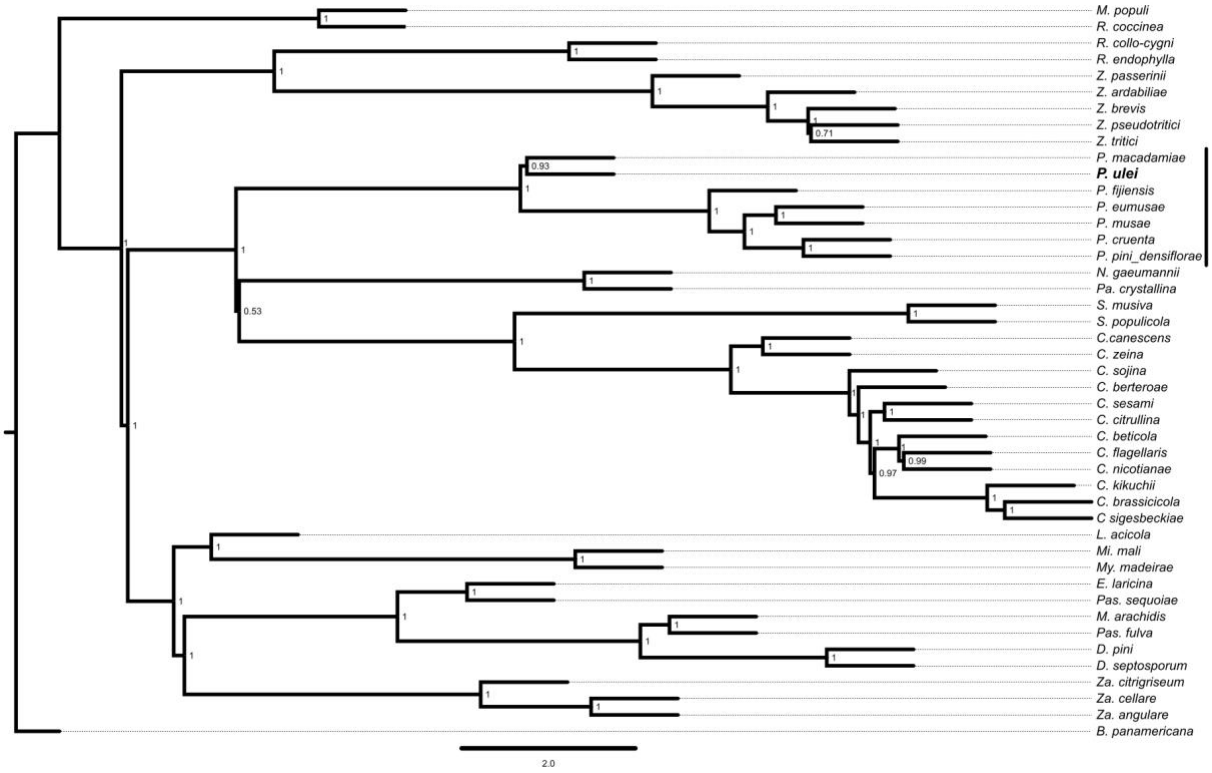


Figure 2-16. Phylogenetic inference of members of Mycosphaerellaceae which genome sequences are available (see Appendix 2 for the genome accession number for each species). *P. ulei* appears in bold face. Members of the genus *Pseudocercospora* are identified by a vertical black line.

2.3.12. Identification of repetitive content

With over 80% of the total genome size, *P. ulei* shows a high transposable element content (Figure 2-17). Most TEs belong to retrotransposons of the superfamily Gypsy, which comprise 46.09% (16’904) and not yet classified TE superfamilies (29.53%/22’129 elements) (Figure 2-18). Regarding the other *Pseudocercospora* species, they exhibited less number and diversity in Class I - LINE subclass (non-LTRs) than the identified in *P. ulei*, represented by Tad1 and L1 superfamilies with except for *P. pini* for which LINEs TEs was not identified (Figure 2-18). Conversely, Class II TEs were more diverse than *P. ulei* for which DNA TEs were distributed only in the harbinger, hAT, and CMC- EnSpm superfamilies (Figure 2-18).

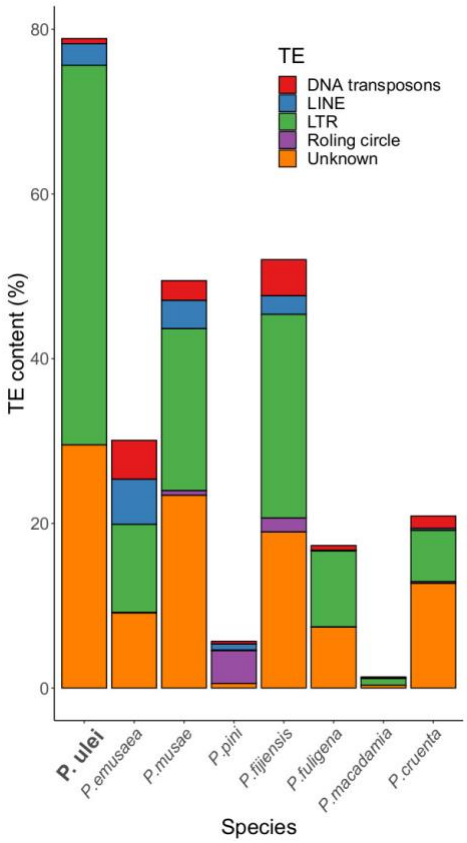


Figure 2-17. general overview of repetitive content in the *Pseudocercospora* genus

It has been described that the genomes size differences among Dothideomycetes fungi class is strongly and positively correlated with transposable element proliferation (Haridas et al., 2020). This relationship is observed in the sequenced genomes of *Pseudocercospora* species. For instance, those that form of the Black Sigatoka complex of banana crops where the largest genomes (*P. fijiensis* and *P. musae*) content 52.7 and 50.1% of transposable elements respectively, while the smallest genomes (*P. macadamia* and *P. pini*) presented fewer transposable elements (Figure 2-17). Our results confirm the impact of transposable elements on genome size expansion, since not only the *P. ulei* is the largest genome so far reported in the Mycospharelaceae family but the genome with the highest TEs content.

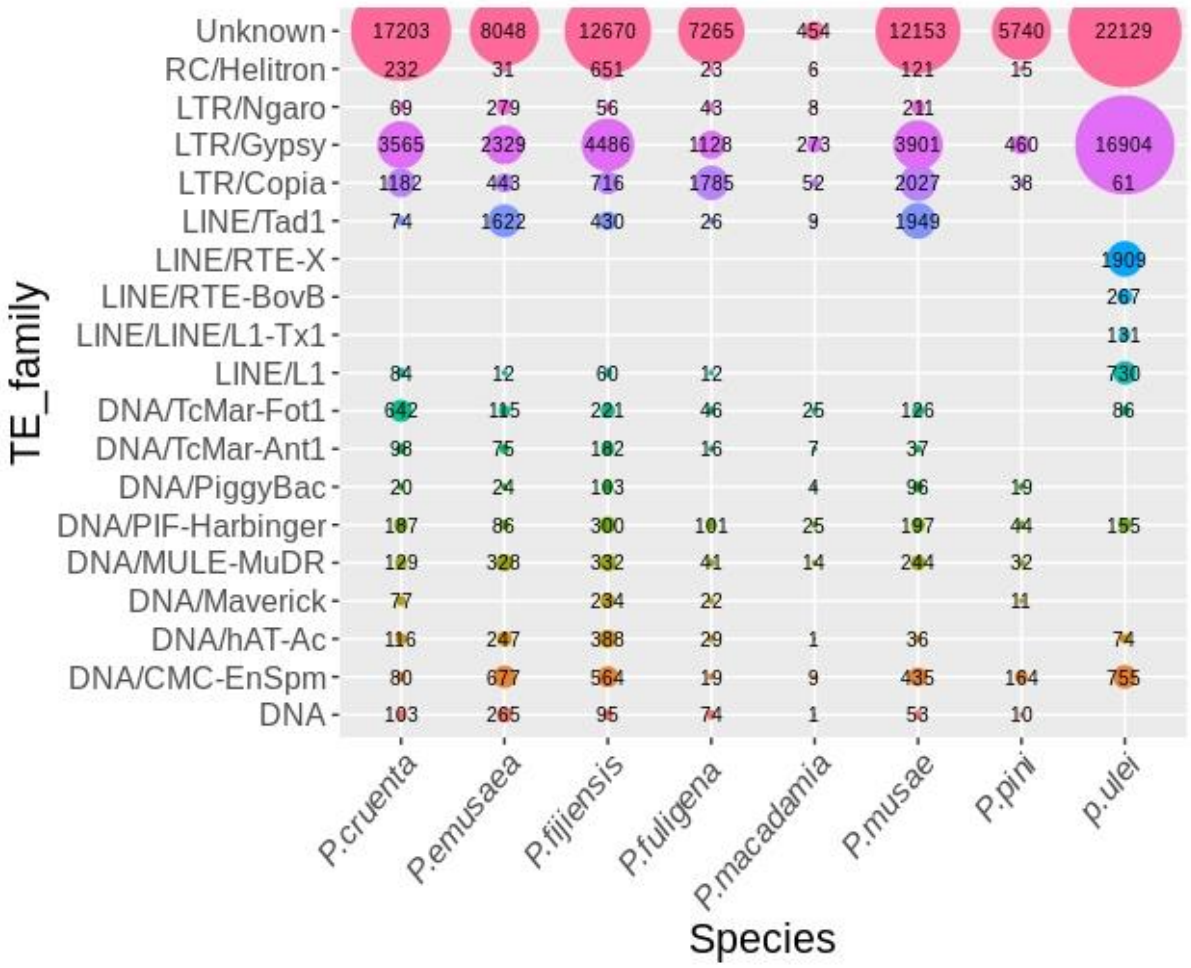


Figure 2-18. Number of copies per TEs families present in *Pseudocercospora* genus sequenced genomes.

2.4. Final remarks

Nowadays, the assembly of a reference genome sequence for a species is critical for most basic and applied research, including the systematic identification of pathogenicity genes. The dramatic reduction in DNA and RNA sequencing costs (Goodwin et al., 2016) and the accessibility to a significant number of fungal genomes since the 1000 genomes project advent (Grigoriev et al., 2014) have revealed the complexity hidden in the fungal genomes. In this study, we sequenced, assembled, and annotated the genome of *P. ulei*, a poorly studied fungus and the most limiting pathogen of natural rubber tree (*Hevea brasiliensis*).

The *P. ulei* genome was sequenced through a hybrid strategy generating reads from the second and third generation sequencing technologies. The best genome assembly was built with PacBio reads by hierarchical methods using Canu. The assembled *P. ulei* genome has 93.8 Mb and was fragmented into 231 contigs, where the largest was 11.2 Mbp, (N50 = 2.35 Mbp), GC% = 50.27%. BUSCO identified 1’282 complete orthologous genes (97.5%) in the assembly. From these, 1’273 were single-copy, and only 9 (0.7%) were duplicated, suggesting that this genome is haploid. The 231 *P. ulei* genome contigs were processed into 215 scaffolds; seventeen scaffolding events increased the N50 value to 2.82 Mbp. The *P. ulei* whole-genome assembly was deposited in GenBank with the accession number JACWNB000000000 under BioProject PRJNA661960.

The *P. ulei* gene models’ prediction were performed with Braker2 training predictors with evidence from several resources, including RNAseq reads from *P. fijiensis* (SRP075820) and *P.ulei* which were generated in this project. Each evidence resource contributed in the gene structure prediction, thus, *Pseudocercopora* protein databases strengthened the exons annotation, and transcripts evidence increase the prediction of exon-intron junctions and splicing sites. The most efficient genes prediction were performed in the *P. ulei* unmasked genome, merging the whole assessed evidence, and allowed us to identified a total of 12’745 protein-coding genes, with a mean length of 2’075 bp and a gene complete BUSCOs content of 91.3%.

The phylogenetic reconstruction enable to stablished that the closest relative of *P. ulei* is *P. camelliicola*, a pathogen of camellia cultivars. The topology based on the four loci markers showed that *P. ulei* was placed into a *Pseudocercospora* species group that has not been genetically explored yet, being the *P. ulei* genome, the first in this group. On the other hand, the orthologous genes phylogenomic analysis confirmed the previous findings, with *P. fijiensis*, *P. eumusae*, *P. musae*, *P. cruenta* and *P. pini_densiflora*, forming a clear clade, and *P. ulei* and *P. macadamia* forming another one, and being sisters to each other. The genome of *P. ulei* contributes new genome information in a group of *Pseudocercospora* that is still not fully explored from the genomics point of view, and could help answer evolutionary and biological unsolved questions for these plant pathogens.

Chapter 3

3. The functionality of *P. ulei* genome.

3.1. Introduction

The combination of a growing number of high quality genome assemblies for different fungi with RNA-seq assays enables the systematic identification of proteins secreted by a microbial agent for pathogenicity or virulence (CAZymes, and Small Secreted Proteins “SSPs”) as well as other molecules implied in the pathogenicity process like Secondary metabolites i.e., toxins. The collection of these proteins is known as Secretome (Ballester et al. 2015; Sperschneider et al., 2016; Sandhya et al., 2016). A comprehensive secretome data makes a useful information resource to build predictive models of gene dynamics driving the infection mechanisms during the pathogen invasion. The secretome has been typically characterized by comparative genomics of the amino acid sequences of genes predicted from genome assemblies. For example, the *U. maydis* and *Fusarium graminearum* secretomes were generated predicting proteins that are secreted by the conventional secretion system via Endoplasmic Reticulum - Golgi Apparatus by searching the signal peptide and translocation motifs (Mueller et al., 2008; Brown et al., 2012). This was achieved mapping the annotated genes against SignalP and TargetP databases and translocation motifs through 3-dimensional structure prediction by PROTCOM database. The *U. maydis* and *Fusarium graminearum* secretomes are constituted by 426 and 574 putative secreted proteins respectively. These results are contrasting with the more than 1,535 proteins constituting the secretome composition of the hemibiotrophic fungus *Moniliophthora roreri* (Meinhardt et al, 2014). The smaller number of genes in the secretomes of the biotrophic fungi *U. maydis* and *F. graminearum* is consistent with the limited hydrolytic activity associated with the biotrophic lifestyle.

The CAZymes and secondary metabolites are part of the pathogenicity associated proteins major studied in fungal models. CAZymes are responsible for the degradation, synthesis, and modification of complex carbohydrates and glycoconjugates (Cantarel et al, 2009). Taking into consideration that the

carbohydrates are one of the main constituent of the plants cell walls, these enzymes key role in the fungal invasion and colonization process, mediating the host physical barriers modification. The number and composition of CAZymes are nowadays the source to the fungal life style prediction (Hane et al., 2020). Carbohydrates are not the only component that fungi face up during ta colonization process, vegetal cell walls are also composed by lipids and proteins which give rigidity to this complex structure. Therefore, the arsenal of enzymes secreted by fungi in the early disease stages also include lipases and peptidases addressed to alter the surface characteristics facilitating fungal adherence and penetration (Pascoal et al., 2018; Krishnan et al., 2018).

The secretome of a fungal species includes the entire set of effector proteins which, as explained above, are expressed at different specific stages of the infection process. Availability of a growing number of fungal genomes allow to identify characteristic motifs and domains in the amino acid sequences of effectors (Govers and Gijzen, 2006), as well as the determination of cysteine-rich residues and signal peptide sequences (Liu et al., 2012). Although most of these signals are not completely conserved across the fungi kingdom, they provide partial cues for identification of effectors in assembled genes. Assembled genomes are also a source of additional partial information on effectors because fungal effectors tend to be located in low gene density compartments, organized in clusters and flanked by repetitive sequences and transposable elements (Dong et al., 2015; Möller and Stukenbrock, 2017). An example of the combination of these partial sources of information is the prediction and further confirmation of the SnTox1 effector of *Stagonospora nodorum* (Liu et al., 2012). The authors predicted effector candidates from *S. nodorum* genome assigning a score to effector candidate genes that accomplish criteria as the presence of signal peptide, cysteine residues, and RXLR or RGD motifs, a predicted size of 30 kDa, gene localization within a repetitive sequence of 5 Kb and gene homology with proteins previously detected by mass spectrophotometry in culture filtered. The SnTox1 effector was identified from the genes with first 100 scores aligning them against the non-redundant protein

database. SnTox1 showed high homology with the effector Avr4 of *C. fulvum* demonstrating that both effectors share structural features like chitin binding domain and six cysteine residues. Finally, localization assays using green fluorescence protein GFP suggested that SnTox1 was present at the apoplastic level. As described in (Sperschneider et al.,2016), modern machine learning approaches could be promising to build more accurate prediction models to identify effector proteins combining the partial sources of information related to semi-conserved domains, residues composition and genome localization (Sperschneider et al, 2018).

Despite significant progress in genome assembly techniques, effectively discovering effector genes remains challenging because of their localization in genomic compartments of low gene density and high repetitive elements content. Effector-rich genome compartments often have higher mutation rates and undergo frequent rearrangements (Fouché, Plissonneau & Croll, 2018). The compartmentalization of the pathogen genome was proposed as the “two-speed genome model” (Plissonneau et al., 2017). The low sequence complexity and the lack of homology make effector gene prediction challenging. This was demonstrated by the AvrStb6 effector gene of *Zymoseptoria tritici* AvrStb6 was missing from genome annotations for a decade (Goodwin et al, 2011) before mapping approaches identified the gene (Zhong et al, 2017). Hence, highly contiguous genome assemblies are critical for *de novo* effector gene prediction.

In this chapter we characterize the predicted gene functions, focusing on the pathogenicity-related proteins like CAZymes, effectors, secondary metabolites, and transposable elements (TEs). Finally, the collected information throughout genome assembly and functional annotation will be adapted into a comprehensive model of the *P. ulei* and *H. brasiliensis* molecular interaction, given for the SALB disease development.

3.2. Materials and methods

3.2.1. Comparative analysis datasets

For comparative functional analysis across the genus, we obtained the previously sequenced *Pseudocercospora* genomes and the annotated protein sequences if they were available from NCBI database (See the accession numbers in Table 12). We used eight *Pseudocercospora* spp genomes available in public databases, comprising i) *P. fijiensis*, *P. musae*, and *P. eumusae* species, the Sigatoka disease complex in the banana crop (Isaza et al., 2016), ii) *P. macadamia*, the causal agent of husk spot in macadamia crop (Akinsanmi & Carvalhais, 2020), iii) *P. cruenta*, the causal pathogen of *Cercospora* leaf spot in cowpea crops, iv) *P. pini-densiflora*, a cosmopolitan pathogen that affects several species of pines, and *P. fuligena*, a tomato pathogen that causes the black leaf mold (Table 1-13).

For the secretome prediction, we used the protein sequences available for the Sigatoka disease complex (Chang et al., 2016), and *P. macadamia* (Akinsanmi & Carvalhais, 2020). For the *Pseudocercospora* species without annotation (*P. pini*, *P. cruenta*, and *P. fuligena*) we performed the gene prediction with BRAKER2 using a protein non-redundant database of the *Pseudocercospora* genus, and subsequently obtained the protein sequences using gffread (Pertea & Pertea., 2020).

Table 1-13. *Pseudocercospora* species used for comparative genome analyses

Species name	Accession number	Genome Size	Number of genes	Disease	Crop	Reference
Pseudocercospora cruenta	SM1336520v1	55,623,252	14.228	Cercospora leaf spot	cowpea	not reference
Pseudocercospora eumusae	ASM157823v1	47,119,461	12.632	eumusae leaf spot	Banana	Chang et al.,2016
Pseudocercospora fuligena	ASM1429803v1	50,525,800	14.494	black leaf mold	Tomato	Zaccaron & Stergiopoulos, 2020
Pseudocercospora fijiensis	Mycfi2	74,141,167	13.066	black sigatoka	Banana	Isaza et al.,2016
Pseudocercospora macadamiae	ASM1297840v1	40,070,143	15.430	husk spot	Macadamia	Akinsanmi & Carvalhais, 2020
Pseudocercospora musae	ASM157822v1	60,439,160	13.129	yellow sigatoka	Banana	Chang et al.,2016
Pseudocercospora pini-densiflorae	ASM50436v2	43,513,371	14.555	Needle blight of pine	Pine	not reference
Pseudocercospora ulei	IBUN_P_ulei_V1	93,730,151	12.745	South American Leaf Blight disease	Rubber	This study

3.2.2. Functional gene annotation

Functional annotations were performed with InterProScan v.5.36-75 switching on the *-goterms* and *-iprlookup*, and -pathway options. In addition, we obtained the general statistics of the functional annotation using the TRAPID program which annotates the assembled transcripts (Van bel et al,.2020).

3.2.2.1. Pathogenicity candidate proteins annotation

We predicted the putative pathogenicity proteins for all *Pseudocercospora* species applying the secretome pipeline shown in Figure 3-1.

3.2.2.1.1. *CAZymes and Secondary Metabolites genes annotation*

The *P. ulei* pathogenicity proteins prediction also encompassed the identification of CAZymes and secondary metabolites. The CAZymes were predicted from the *P. ulei* proteins using dbCAN2 Last update 09/08/19 (Yin et al., 2014), The results of the *P. ulei* CAZymes composition were used to inferred *P. ulei* lifestyle using the CATAStrophy V0.1.0 (Hane et al., 2020). The gene clusters for the synthesis of secondary metabolites were identified from *P. ulei* genome assembly and the structural annotation General Feature Format (GFF) using antiSMASH v.5.0 (Blin et al., 2013) (Figure 3-1).

3.2.2.1.2. *Effectors prediction*

To annotate effectors, peptidases and lipases candidates, we created a catalogue of predicted secreted proteins from the total of coding regions predicted. For that end, we selected the proteins with signal peptides using SignalP v.5.0b and Phobious November 20/2020 – online version (Käll et al., 2007; Nielsen et al., 2017), and then we identified and removed proteins with: transmembrane domains, signals targeting endoplasmic reticulum, subcellular location signals and GPI anchors using TMHMM2 (Viklund & Elofsson. 2004), SCAN_prosite (Castro et al., 2006), WolfPSORT v. 0.2 (Horton et al., 2007), and Pred-GPI (Pierleoni, Martelli &

Casadi. 2008), respectively. Putative effectors were determined with effectorP V2.0 software (Sperschneider et al., 2015), and the localization was predicted using ApoplastP (Sperschneider et al., 2018).

Secreted lipases and peptidases were identified aligning the secreted proteins catalog against the Lipase Engineering Database (http://www.led.uni-stuttgart.de/) and the MEROPS blast database (Rawlings et al., 2014) respectively. The alignments were performed with BlastP using a similarity of >=80% as an inclusion parameter. Non secreted lipases and peptidases were predicted from the whole *P. ulei* proteome following the described steps.

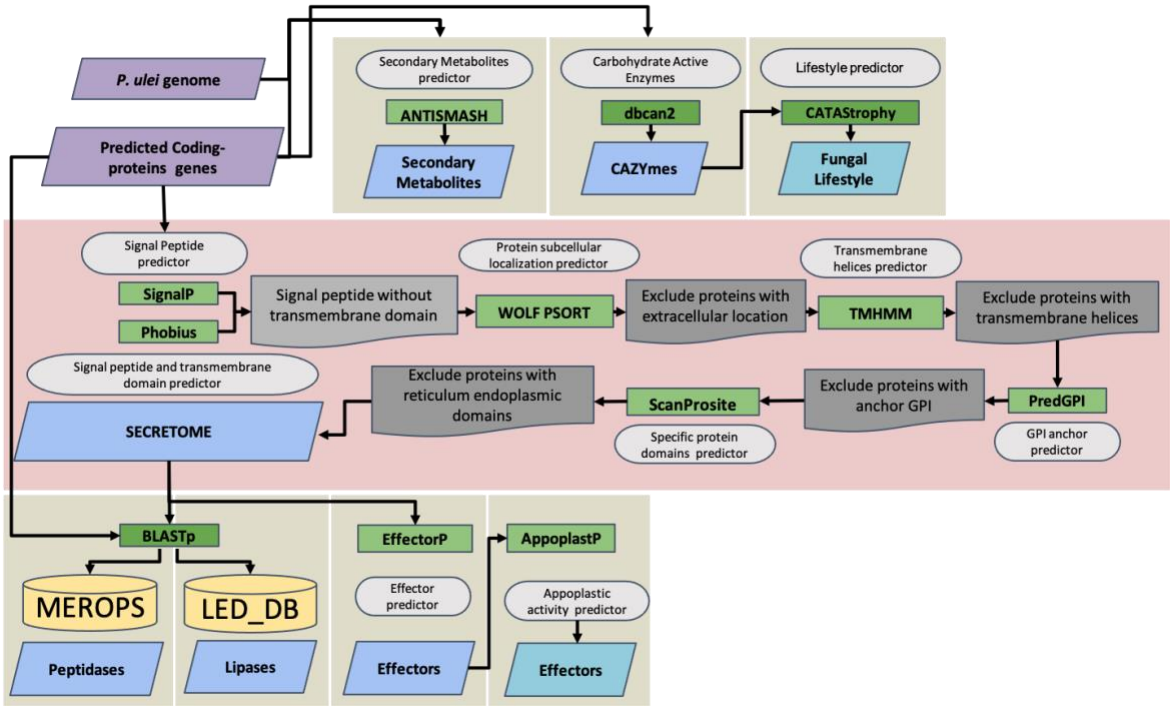


Figure 3-1. Pathogenicity proteins prediction pipeline

3.2.2.2. Transposable element annotation

We used a set of previously published TE family consensus sequences from eight *Pseudocercospora* spp genomes, that was created with RepeatModeler v1.0.11 (http://repeatmasker.org), based on sequences in the GIRI Repbase (González et al, 2021; Bao et al, 2015). To remove redundancy, fragmented consensus sequences and to gain better precision on the sequences, we did a manual curation

of the existing consensus sequences, using the WICKERsoft pipeline (Breen et al, 2007). This pipeline is extracting copies of the same TE family with 300bp added both up-and downstream, and is creating multiple sequence alignments. From these alignments we manually defined the start and end site. In comparing the sequences with dot plots, we removed redundancy. Additionally we screened the genome for non-autonomous TE families, using LTR-Finder to detect LARDs (LArge Retrotransposon Derivates) and TRIMs (Terminal Repeat retrotransposons In Miniature), MITE-Tracker to detect MITEs (Miniature Inverted-repeat Transposable Elements) and SINE-Finder in SINE-can to detect SINEs (Short Interspersed Nuclear Elements) (Xu & Wang, 2007; Gao et al, 2016; Ma et al, 2015; Crescente et al, 2018; Mao & Wang, 2017). With the obtained TE consensus sequences, we run RepeatMasker v.4.0 to detect all occurrences in the genome (Smit et al, 2015).

3.3. Results and Discussion

3.3.1. Functional annotation

InterproScan assigned PFAM domains to 60.5% (n= 7’721) of the 12’745 gene models with the most abundant categories being RNase H-like domains found in reverse transcriptase (17,8%), reverse transcriptase (RNA-dependent DNA polymerase) (13.3%), integrase zinc binding domain (8.4%), WD domain, G-beta repeat, and protein kinase domain (0.8%) (Figure 3-2). A total of 9.8% (n= 1’254) of the functionally annotated genes presented gene ontology terms (GO) and 15.4% (n=1’975) were associated with a biological function registered in the KEGG database.

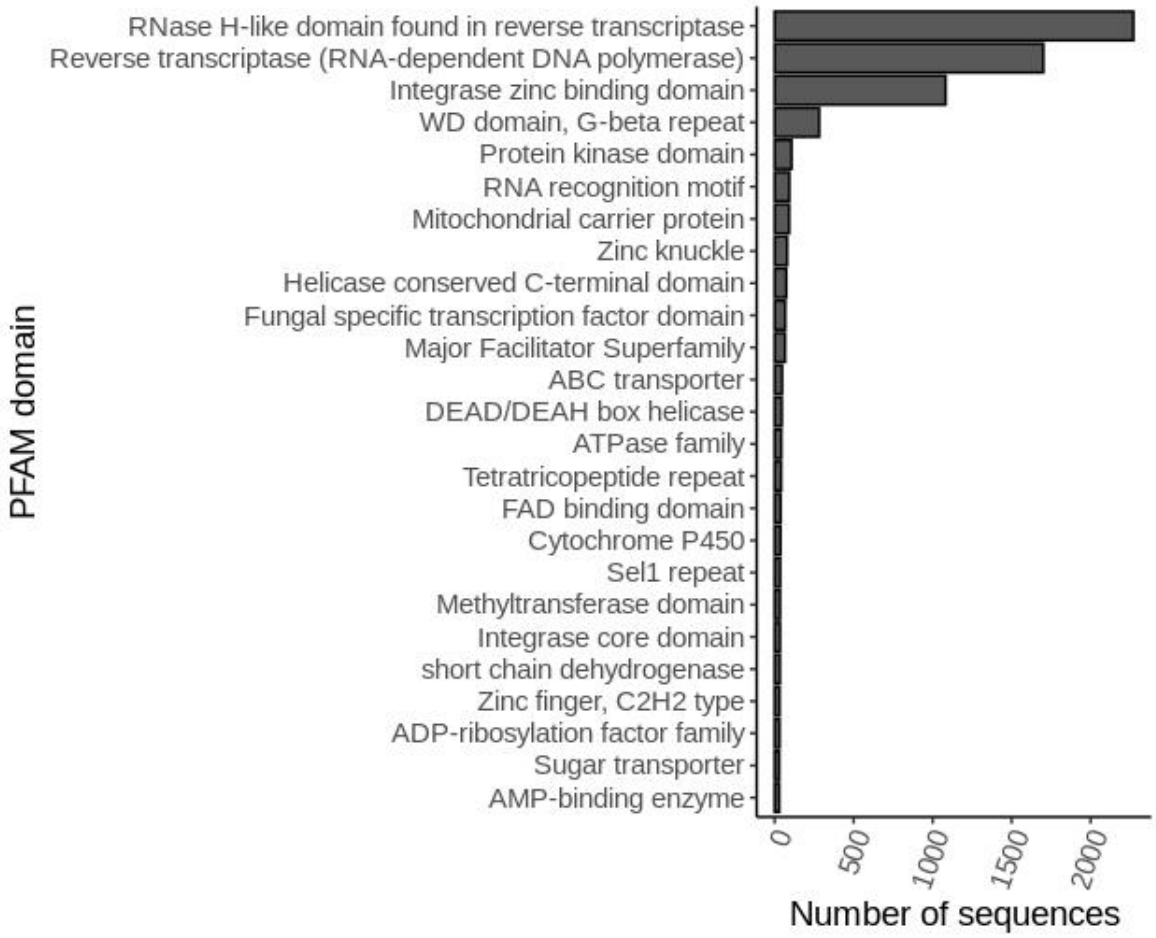


Figure 3-2. PFAM families assigned to the *P. ulei* genome.

The most abundant GO assignments were in the molecular function category, with the most prevalent being those associated with protein binding (n=3’206), nucleic acid binding (n=1’576), catalytic activity (n=482), iron ion binding oxidation process (n=431), ATP binding (n=362), and protein kinase (n=390) activities (Figure 3-3). The biological process categories involved mainly phosphorylation as well as cellular components were mostly associated with nucleotide binding and protein transport and integral components of membrane (Figure 3-3).

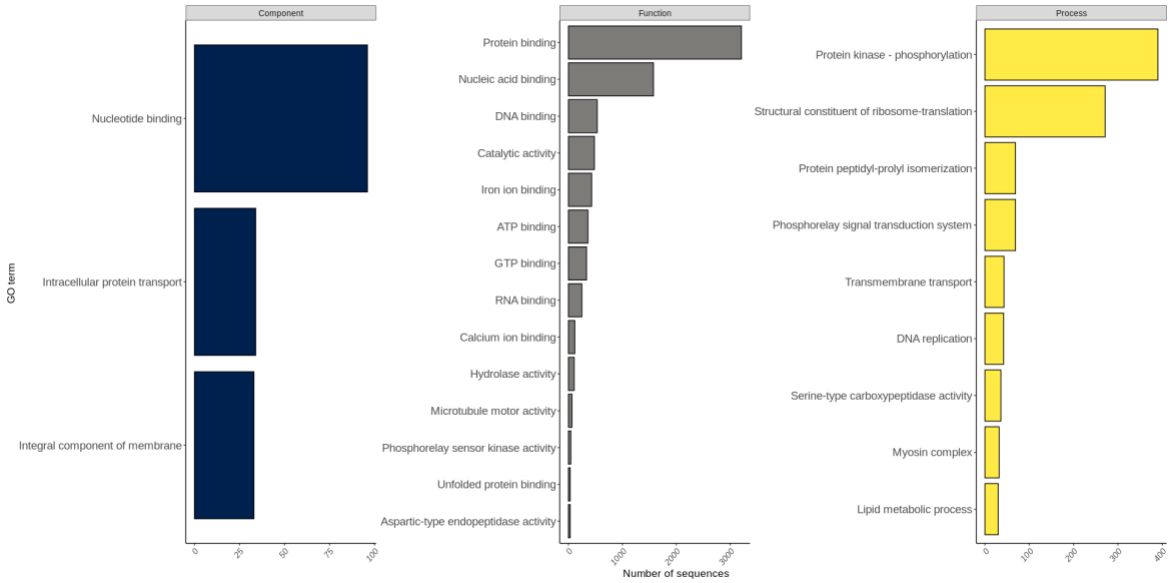


Figure 3-3. GO classification of *P. ulei* proteins.

The transcripts analysis with TRAPID revealed that 9'250 (72.6%) were assigned to 5'265 gene families with the largest being 2724_0PGPZ, which corresponds to the categories reverse transcriptase from replication, recombination and repair category. Regarding the taxonomic classification 11'345 (89.02%) of transcripts were classified while 1'400 (10.98) remained unclassified. 9'031 (70.86%) of the classified transcripts were assigned to eukaryote domain of which 5'432 corresponded to the Dothideomycetes class, 5'066 to the Capnodiales order and 4'671 to the Mycosphaarelaceae family which contain the *Pseudocercospora* genus, represented with 3'706 transcripts.

3.3.2. Pathogenicity gene candidates

Approximately 5.7% (n=756) of *P. ulei* coding sequences were classified as potentially secreted proteins. Among them, 1.6% (n=216) were predicted as putative CAZymes, 0.8% (n=113) as putative effectors, and 0.42% (n=54) presented a potential activity in the host apoplast (Table 1-14). 91 protein-coding genes were annotated as putative peptidases, however, only 15 of those had signal peptides (Table 1-14). The latest is a low number compared to the other *Pseudocercospora* species which had a rank of 51 and 107 of potentially secreted

peptidases (Table 1-14). Likewise, lipid degrading activity was poorly represented in the *P. ulei* genome with only one gene being annotated to lipase enzymes (Table 1-14).

The secretome was considerably smaller in *P. ulei* compared to the other analyzed *Pseudocercospora* species (Figure 3-4). The hemibiotrophic pathogen *P. eumusae* presented the largest proportion of putatively secreted proteins, with 21.94% (n=2'898), followed by the hemibiotrophic *P. pini* with 21.55% (n=3'136) and the necrotrophic *P. macadamia* with 14.54% (n=2'230). The secretome profiles for all *Pseudocercospora* species were primarily consisting of CAZymes and putative effectors (Figure 3-4). Inversely, peptidases and lipases were poorly represented in the *Pseudocercospora* spp genomes compared to the necrotrophic *Corynespora cassiicola,* a broad range pathogen that affects more than 500 plant species (Lopez et al., 2018).

Table 1-14. Overview of pathogenicity candidates annotated in the *Pseudocercospora* genus species.

Secretome	*P.cruenta*	*P.emusea*	*P.fijiensis*	*P.fuligena*	*P.macadamia*	*P.musae*	*P.pini*	*P.ulei*
Total genes	14228	12632	13066	14494	15430	13129	14555	12745
Secreted proteins	1205	2898	1128	1926	2230	1910	3136	756
CAZYmes	475	403	374	492	490	396	489	216
Secondary metabolite clusters	44	28	30	35	36	27	36	14
Effector	160	423	195	241	379	312	453	113
Appoplast activity	97	219	84	105	190	121	246	54
Lipases (Not secreted)	11	6	4	6	4	2	0	1
Lipases (Secreted)	5	5	5	5	4	4	13	0
Peptidases (Not secreted)	146	87	154	139	105	116	96	91
Peptidases (Secreted)	52	89	51	65	43	62	107	15

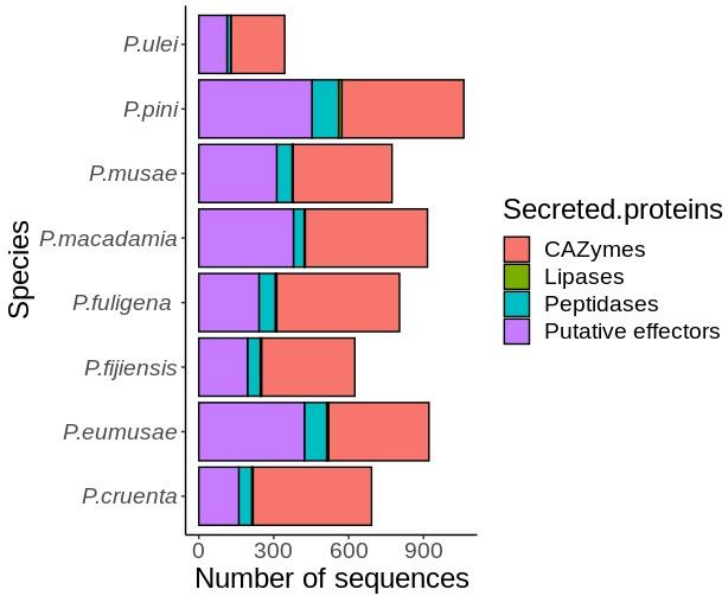


Figure 3-4. Secreted proteins profile of the sequenced *Pseudocercospora spp* genomes.

3.3.3. Carbohydrate Active Enzymes

The 216 proteins annotated by dbcan2 were classified into 96 CAZymes families. From them, glycoside hydrolases and glycosyltransferases were the most abundant families representing 47.6% (n=103) and 39.8% (n=86), respectively, followed by Carbohydrate-binding module (CBM) and auxiliary activities (AA) which formed each 5% (n=11) and 6% (n=13). Whereas the polysaccharides lyases (PL) and carbohydrate esterases (CE) families were underrepresented with 0.96% (n=2) and 0.46% (n=1), respectively (Figure 3-5).

The number of CAZymes in *P. ulei* genome is among the lowest in a set of 69 mycelial ascomycetes (Hane et al., 2020), and in the seven *Pseudocercospora spp*. sequenced genomes (Figure 3-5). The low proportion of CAZyme content has been described as an evolutionary trait of hemibiotrophic and biotrophic lifestyles within the Capnodiales (Haridas et al., 2020). Moreover, CATAStrophy, a lifestyle predictor that uses the CAZymes profile, indicated that *P. ulei* has a monomerthroph lifestyle (relative centroid distance = 1), which is roughly analogous to a biotrophic lifestyle characterized primarily by simple sugars metabolism (Hane et al.,2020).

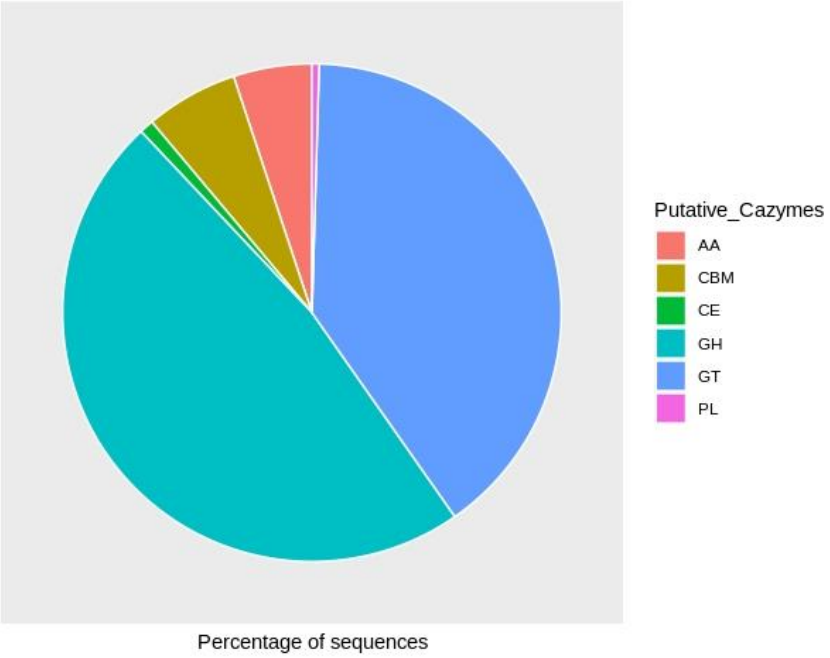


Figure 3-5. CAZymes profile of *P. ulei* genome. AA corresponds to auxiliary activities, CBM to Carbohydrate-binding module, CE to carbohydrate esterases, GH to glycoside hydrolases, GT to glycosyltransferases, and PL polysaccharides lyases.

An in depth analysis of the *P. ulei* CAZyme composition revealed reduced pectin degrading activity, as identified by the absence of PL1, PL3, PL9, GH28 families. Moreover, a limited assignation of PL4, GH78, CMB50, and GH105 families was found, with only one protein-coding gene assigned per family. Likewise, the hemicellulose degrading activity was poorly represented in *P. ulei* compared with the other analyzed *Pseudocercospora* species (Figure 3-6), which were better represented in the CE1, CE4, GH3, GH16, GH43, and GH51 families. Among CAZymes associated with cellulose degradation, *P. ulei* had between one and seven proteins assigned to GH5, GH6, and GH7 families, respectively. Conversely, *P. ulei* cellulose degrading proteins assignments were similar to those obtained for the necrotrophic *P. macadamia,* with between one and ten predicated CAZymes.

The CAZymes results reflect that the *P. ulei* infection process could be defined by a minimized plant cell wall degradation which might be focused on cellulose degradation. Similar results were reported for the rust fungi *Melampsora larici-populina* and *Puccinia graminis f. sp. tritici*, describing the active role of the GH5 CAZyme family genes (*β*-1,3-Glucanases and Endo-1,4 *β*- glucanase activities) in the early infection process. The authors highlighted the essential function of these genes which soften the parenchyma cells before the host penetration (Duplessi et al., 2011).

Albeit the differences in nutritional lifestyle, our data suggest that *Pseudocercospora* species could have a phylum-specific CAZymes distribution (Figure 3-6). The *Pseudocercospora* species CAZymes composition is constituted in descending order by GH, GT, AA, CMB, CE, and PL families (Figure 3-6), being these latest represented by no more than nine genes. This is a low proportion, especially for the CE family which is one of the more abundant in plant fungal pathogens (Zhao, 2013, Lelwala et al., 2019; Rajarammohan et al., 2019). Contrary, the carbohydrate metabolism within the genus *Pseudocercospora* is largely dominated by glycoside hydrolases (GH) and glycosyltransferases (GT) activities (Figure 32).

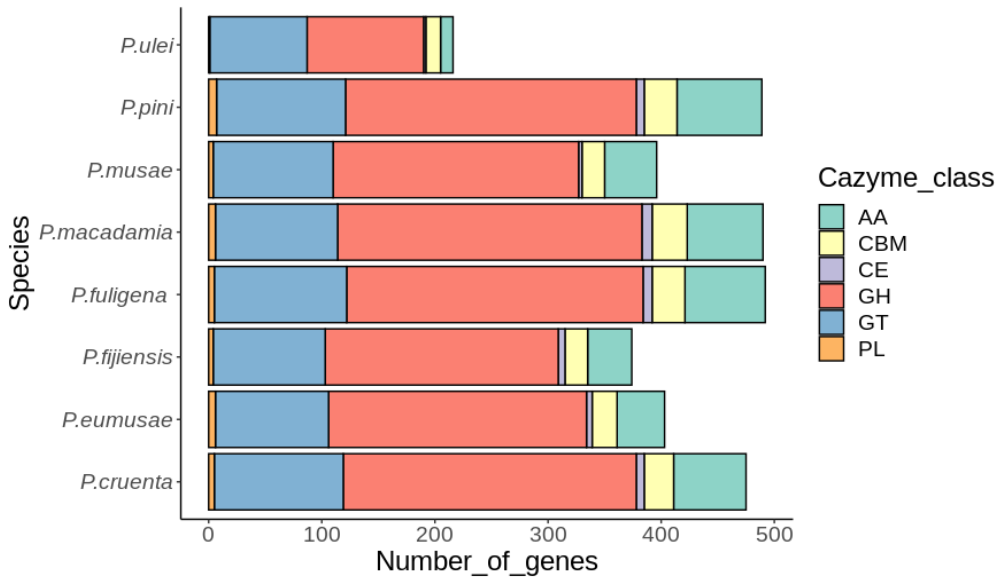


Figure 3-6. CAZymes profile of the sequenced *Pseudocercospora* spp genomes. AA corresponds to auxiliary activities, CBM to Carbohydrate-binding module, CE to carbohydrate esterases, GH to glycoside hydrolases, GT to glycosyltransferases, and PL polysaccharides lyases

3.3.4. Secondary metabolites

The *P. ulei* genome contains a low number of gene clusters associated with secondary metabolites (SMs). A total of fourteen SMs backbones genes were identified in the *P. ulei* genome. NRPS (n=3), NRPS-like (n=5), and terpenes (n=3) were the most abundant. Additionally, *P. ulei* genome has two T1PKs, which are generally associated with the fungal toxin synthesis (Kroken et al., 2003). One of the T1PKs gene clusters identified in *P. ulei*, presented 100% of similarity with a melanin synthesis gene cluster. Melanin is a phenolic pigment which is considered as a pathogenicity factor due to its role in the appressorium formation (Jacobson 2000; Langfelder et al., 2003; Scharf et al., 2014) and resistance against host (Casadevall, Rosas & Nosanchuk, 2000). These results support previous findings by Hashim, Cheen, and Duncan (1978), who determined through fluorescence observations that the *P. ulei* invasion process is likely initiated by appressorium formation.

The secondary metabolites gene clusters were considerably smaller in *P. ulei* compared to the other analyzed *Pseudocercospora* species (Figure 3-7) that coded among 27 - 44 gene clusters (Table 1-14). The comparative results across the genus showed that *P. ulei* only had half as many putative SM gene clusters as other species in the black Sigatoka complex, and 2.5 times fewer than the necrotroph *P. macadamia* (Table 1-14). The low number SMs genes in *P. ulei* also confirms its biotrophic lifestyle. The convergent loss of these genes is because, biotrophic lifestyle often implies a fewer activity of secondary metabolites enzymes, as well as a lower toxins secretion compared to necrotrophs fungi (Pusztahelyi, Holb & Pócsi, 2015).

In accordance with their nutritional metabolism, the seven *Pseudocercospora* species showed a higher diversity in SM gene clusters than *P. ulei.* Notably, genes

cluster assigned to enzymes required for siderophore biosynthesis were only found in the seven *Pseudocercospora* species, while absent in *P. ulei* (Figure 3-7). Siderophores act in the iron uptake and homeostasis, iron is an essential element for biomolecules anabolism, and cell-damaging hydroxyl radicals production (Haas, Eisendle & Turgeon, 2008). Due to the importance of iron in the survival of eukaryotic microorganisms, siderophores are considered as fungal virulence factors for their role in iron kidnapping from the host tissue (Scharf et al., 2014).

The absence of siderophore biosynthesis enzymes in *P. ulei* might suggest that iron uptakes could be mediated by an alternative system than siderophore-assisted iron uptake. Reductive Iron Assimilation (RIA) is the second high affinity system employed by fungi for iron uptake (Johnson, 2008). RIA Iron uptakes consists in a extracellular reduction of Fe3+ to Fe2+ by transmembrane ferrireductases, Fe2+ is subsequently re-oxidated and imported by a bipartite protein complex comprehended by a ferroxidase (FetC) and an iron permease (FtrA) (Blatzer, Binder & Haas, 2011). It is highly plausible that RIA be the central system for *P. ulei* iron ions assimilation. This can also be corroborated with the functional annotation, which identified 431 assignments to GO:0005506 term that corresponds to the iron ion binding, being the fifth more abundant function found in *P. ulei*. (Figure 3-7).

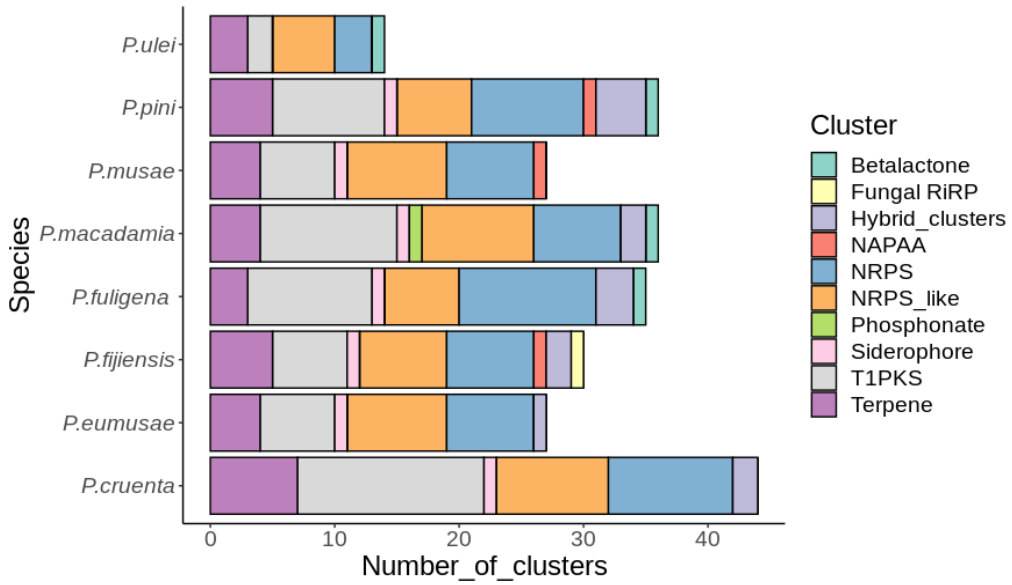


Figure 3-7. Secondary metabolism gene clusters present in *Pseudocercospora spp* genomes. NRPS correspond to Non-Ribosomal Peptide Synthetase, and (PKS) to Polyketide Synthase.

3.3.5. *P. ulei* predicted effectors

A total of 113 secreted protein candidates were classified as putative effectors by the EffectorP program (Table 1-15). Of those, 54 putative effectors presented potential activity in the host apoplast, while and 59 were annotated as effectors with activities of unknown host location (Table 15). The length of effector candidates varies between 100 and 358 amino acids. The tblasn assignments revealed that 67% (n = 92) of effectors candidates presented similarity with proteins belonging to species belonging to the family of Mycosphaerellaceae. From them, 18% [UO1] presented similarity to *Cercospora beticola* proteins, 11% with *Pseudocercospora fuligena* and *Ramularia collo-cygni*, 10% with *Pseudocercospora fijiensis*, 8% with *Sphaerulina musiva*, 7% with *Pseudocercospora eumusae*, 4% with *Pseudocercospora musae*, and 3% with *Passalora fuljina* and *Cercospora spp*. 24% of the effector candidates were assigned to protein sequences from six fungi belonging to Dothideomycetes, Eurotiomycetes and Zygomycota orders. Meanwhile, 11% (n = 13) of effectors candidates did not show similarity to any sequence in the NCBI databases.

Of the 59 effector candidates without identified location, 8 were assigned to transport functions such as exocytosis. The exocytosis process is widely conserved in plants, animals and fungi, and it has recently been associated with morphogenesis and pathogenicity processes (Chen et al. , 2015). In filamentous fungi, eight genes represented by Sec 3, Sec 5, Sec6, Sec8, Sec 10, Sec 15, Exo70 and Exo84 have been reported to form the exocytosis complex, while paralogous genes have been described in plants (Schultzhaus & Shaw, 2015).

We found Sec15, a RAB type GTPase protein that could correspond to sec4 in *P. ulei* putative effectors catalog (TerBush & Novick,1995; TerBush, Roth & Novick, 1996). Furthermore, we found an effector candidate encoding a SNARE protein

which are subcellularly located and could interact with other exocytosis complex proteins, mediating membrane fusion processes (Han et al., 2017). A protein with homology to NCAP-1 was also found. NCAP-1 is a negative regulator of the AP2 clathrin complex, which has been described as mediator of endocytosis in filamentous fungi (Beacham et al., 2018). The endocytosis and exocytosis protein machinery modulation in filamentous fungi is not only determines the hyphal spread within the host tissue but are also involved in pathogenicity. The exocytosis complex role mediating the cytoplasmic effectors secretion has been described in the rice pathogen *Magnaporthe oryzae* (Giraldo et al.,2013). Deletions in Sec5 and Exo84 exocytosis protein complex decreased the virulence of *M. oryzae* strains, affecting the secretion of PWL2 effectors during the infection process. Conversely, *M. oryzae* apoplastic effectors release was not affected since they are usually secreted by the conventional vesicular pathway via endoplasmic reticulum (Giraldo et al., 2013).

Additionally, we identified an effector candidate with dienelactone hydrolase, a protein with the capability to hydrolyze aromatic-type compounds (Pathak et al., 1988). We hypothesize that *P. ulei* could use this hydrolytic protein to degrade scopoletin, a hydroxycoumarin produced by *H. brasiliensis* whose rapid production and accumulation has been associated with the resistance traits against *P. ulei* (García et al, 1995).

Table 1 -15. *P. ulei* effector candidates without identified location.

N°	ID	Best blast hit	Description	E-value	Size
1	jg1603.t1	XM_033601596.1	Adaptin ear-binding coat-associated protein 1 NECAP-1 (K489DRAFT_322084), partial mRNA [*Dissoconium aciculare* CBS 342.82]	2,00E-66	186
2	jg9136.t1	XM_007927747.1	Carbohydrate esterase family 1 protein (MYCFIDRAFT_70811), mRN [*Pseudocercospora fijiensis* CIRAD86]	8,00E-145	328
3	jg1860.t1	XM_007923870.2	Mannose-6-phosphate receptor CI-MPR (MYCFIDRAFT_120272), partial mRNA [*Pseudocercospora fijiensis* CIRAD86]	3,00E-123	306
4	jg9508.t1	XM_016904683.1	CGI-121-domain-containing protein (SEPMUDRAFT_148073), mRNA [*Sphaerulina musiva*] SO2202	1,00E-61	160
5	jg8101.t1	KX943040.1	Extracellular protein 22 (Ecp22) gene, complete cds [*Passalora fulva* strain 0WU]	3,00E-65	175
6	jg6455.t1	XM_016910441.1	Exocyst complex component Sec15 (SEPMUDRAFT_87355), mRNA [*Sphaerulina musiva* SO2202]	3,00E-64	128
7	jg207.t1	XM_023604348.1	FAD-linked sulfhydryl oxidase ERV2 (CB0940_11544), mRNA [*Cercospora beticola*]	2,00E-74	226
8	jg8984.t1	XM_007927721.1	Glycosyltransferase family 1 protein (MYCFIDRAFT_61) [*Pseudocercospora fijiensis* CIRAD86]	9,00E-89	196
9	jg11567.t1	XM_023766142.1	GTPase Rab7 (RCC_00926), partial mRNA [*Ramularia collo-cygni*]	3,00E-140	205
10	jg4413.t1	XM_016904415.1	LexA/Signal peptidase (SEPMUDRAFT_147614), mRNA [*Sphaerulina musiva* SO2202]	3,00E-48	161
11	jg12468.t1	XM_016907676.1	NADH-quinone oxidoreductase (SEPMUDRAFT_156479), mRNA [*Sphaerulina musiva* SO2202]	3,00E-117	212

12	jg8526.t1	XM_023596849.1	Phosphatidylglycerol/phosphatidylinositol transfer protein (CB0940_04146), mRNA [*Cercospora beticola*]	8,00E-71	186
13	jg290.t1	XM_016909161.1	PR-1-like protein (SEPMUDRAFT_50403), partial mRNA [*Sphaerulina musiva* SO2202]	6,00E-61	249
14	jg11331.t1	XM_003833449.1	Predicted protein (LEMA_P062580.1), partial mRNA [*Leptosphaeria maculans*JN3]	5,00E-15	121
15	jg5407.t1	XM_023767142.1	Probable C-8 sterol isomerase (RCC_01857), partial mRNA [*Ramularia collo-cygni*]	8,00E-123	219
16	jg8249.t1	XM_023773744.1	Probable prohibitin (RCC_08494), partial mRNA [*Ramularia collo-cygni*]	2,00E-176	282
17	jg12122.t1	XM_023765803.1	Probable iron sulfur cluster assembly protein 1, mitochondrial precursor (RCC_00647), partial mRNA [*Ramularia collo-cygni*]	7,00E-106	201
18	jg86.t1	XM_023771822.1	Probable oxidoreductase, short chain dehydrogenase/reductase family (RCC_0655) [*Ramularia collo-cygni*]	4,00E-138	243
19	jg1477.t1	XM_023771822.1	Probable oxidoreductase, short chain dehydrogenase/reductase family (RCC_06559) [*Ramularia collo-cygni*]	5,00E-117	220
20	jg1473.t1	XM_023771822.1	Probable oxidoreductase, short chain dehydrogenase/reductase family (RCC_06559), partial mRNA [*Ramularia collo-cygni*]	4,00E-138	243
21	jg5598.t1	XM_023597819.1	Protein ERP1 (CB0940_05100), mRNA [*Cercospora beticola*]	5,00E-92	223
22	jg11150.t1	XM_023770358.1	Related to dienelactone hydrolase (RCC_05082), partial mRNA [*Ramularia collo-cygni*]	2,00E-64	213
23	jg2854.t1	XM_023766204.1	Related to mitochondrial carrier protein (RCC_00980), partial mRNA [*Ramularia collo-cygni*]	7,00E-115	225

24	jg4408.t1	XM_023771003.1	Related to Phosphatidyl Glycerophosphate GEP4, mitochondrial (RCC_05737), partial mRNA [*Ramularia collo-cygni*]	1,00E-104	185
25	jg3955.t1	XM_023769773.1	Related to ubiquitin (RCC_04495), partial mRNA [*Ramularia collo-cygni*]	2,00E-78	172
26	jg11061.t1	XM_018328760.1	Retrovirus polyprotein (VFPFJ_11693), partial mRNA [*Purpureocillium lilacinum*]	3,00E-28	116
27	jg336.t1	XM_033603551.1	Snare-like protein (K489DRAFT_374742), partial mRNA [*Dissoconium aciculare* CBS 342.82]	8,00E-89	213
28	jg8128.t1	XM_023593659.1	Trafficking protein particle complex subunit 4 (CB0940_01018), partial mRNA [*Cercospora beticola*]	5,00E-90	146
29	jg11953.t1	CP036215.1	Chromosome I, complete sequence [*Cercospora sojina* strain RACE15]	6,00E-43	316
30	jg9431.t1	CP036217.1	Chromosome IX, complete sequence [*Cercospora sojina* strain RACE15]	2,00E-11	209
31	jg7487.t1	CP036209.1	Chromosome VIII, complete sequence, [*Cercospora sojina* strain RACE15]	3,00E-12	208
32	jg11774.t1	CP045653.1	Chromosome I, complete sequence [*Talaromyces marneffei* isolate 11CN-20-091]	1,00E-25	151
33	jg3226.t1	CP036217.1	Chromosome IX, complete sequence [*Cercospora sojina* strain RACE15]		132
34	jg3679.t1	XM_659312.1	Hypothetical protein (AN6800.2), partial mRNA [*Aspergillus nidulans* FGSC A4]	7,00E-15	128
35	jg1666.t1	XM_659312.1	Hypothetical protein (AN6800.2), partial mRNA [*Aspergillus nidulans* FGSC A4]	6,00E-20	104
36	jg4691.t1	XM_023597422.1	Hypothetical protein (CB0940_04710), partial mRNA [*Cercospora beticola*]	5,00E-38	229
37	jg393.t1	XM_023602674.1	Hypothetical protein (CB0940_09892), mRNA [*Cercospora beticola*]	7,00E-58	194
38	jg1840.t1	XP_023454872.1	Hypothetical protein CB0940_10574 [*Cercospora beticola*]	3,00E-08	171
39	jg861.t1	XM_654539.1	Hypothetical protein (AN2027.2), partial mRNA, [*Aspergillus nidulans* FGSC A4]	2,00E-23	150

40	jg3606.t1	XM_023604950.1	Hypothetical protein (CB0940_12139), partial mRNA [*Cercospora beticola*]	3,00E-22	144
41	jg1346.t1	XM_007922470.1	Uncharacterized protein (MYCFIDRAFT_86210), mRNA [*Pseudocercospora fijiensis* CIRAD86]	9,00E-55	134
42	jg3594.t1	XM_007928689.1	Uncharacterized protein (MYCFIDRAFT_175138), partial mRNA [*Pseudocercospora fijiensis* CIRAD86]	1,00E-22	118
43	jg5071.t1	XM_007930904.1	Uncharacterized protein (MYCFIDRAFT_211936), mRNA [*Pseudocercospora fijiensis* CIRAD86]	8,00E-143	257
44	jg1168.t1	XM_007931263.1	Uncharacterized protein (MYCFIDRAFT_212079), mRNA [*Pseudocercospora fijiensis* CIRAD86]	3,00E-31	229
45	jg10740.t1	XM_007928857.1	Uncharacterized protein (MYCFIDRAFT_50439), mRNA [*Pseudocercospora fijiensis* CIRAD86]	5,00E-115	215
46	jg1448.t1	XM_007922246.1	Uncharacterized protein (MYCFIDRAFT_76148), partial mRNA [*Pseudocercospora fijiensis* CIRAD86]	0.037	142
47	jg12482.t1	XM_007930386.1	Uncharacterized protein (MYCFIDRAFT_81433), mRNA, [*Pseudocercospora fijiensis* CIRAD86]	2,00E-43	151
48	jg10968.t1	XM_007927844.1	Uncharacterized protein (MYCFIDRAFT_81781), partial mRNA [*Pseudocercospora fijiensis* CIRAD86]	2,00E-63	216
49	jg4361.t1	XM_007924705.1	Uncharacterized protein (MYCFIDRAFT_83694), partial mRNA [*Pseudocercospora fijiensis* CIRAD86]	9,00E-51	109
50	jg12383.t1	XM_003851669.1	Uncharacterized protein (MYCGRDRAFT_94117), partial mRNA [*Zymoseptoria tritici* IPO323]	2,00E-40	131
51	jg4744.t1	XM_016901128.1	Uncharacterized protein (SEPMUDRAFT_115097), partial mRNA [*Sphaerulina musiva* SO2202]	0.026	90

52	jg4740.t1	XP_007932459.1	Uncharacterized protein MYCFIDRAFT_9423 [*Pseudocercospora fijiensis* CIRAD86]	9,00E-43	176
53	jg6786.t1	XM_007924751.1	Uncharacterized protein (MYCFIDRAFT_181642), mRNA [*Pseudocercospora fijiensis* CIRAD86]	1,00E-30	167
54	jg6790.t1		without assignments		168
55	jg7753.t1		without assignments		140
56	jg4714.t1		without assignments		345
57	jg4452.t1		without assignments		85
58	jg9765.t1		without assignments		102
59	jg8987.t1		without assignments		104

3.3.6. *P. ulei* apoplastic predicted effectors

From the 54 apoplastic effectors predicted in the *P. ulei* genome, only 12% (n = 6) were associated with fungal pathogenicity in other species (Table 1-16). Of those we identified a cutinase 3, a superoxide dismutase [Mn], an acetyl xylan esterase 2, a glycosyl hydrolase of family 61, and a protein with similarity to *P. musae* Avr4 effector (Chang et al., 2016). Additionally, three peptides with similarity to *Passalora fulva* Ecp5, Ecp-20-2, and Ecp-22 effectors were predicted by EffectorP (Table 1-15) but not by ApoplastP.

The role of cutinases in the degradation of host physical barriers has been widely described in fungal phytopathogens during early infection stages of (Gui et al., 2017). Multiple genes from different cutinases families have been identified in *Magnaporthe grisea*, *Fusarium graminearum*, *Botrytis cinerea*, *Aspergillus nidulans*, *Neurospora crassa*, and *Verticillium dahliae* phytopathogens fungal species (Skamnioti et al., 2008). The expansion of these protein families have been associated with the adaptation to infecting multiple hosts (Zhao et al., 2013). Interesting, we identified a single cutinase (PuCut1) in the *P. ulei* effector catalog. The role of the CtCut1 cutinase during fungal adhesion and penetration processes was previously demonstrated through knockout assay in *Colletotrichum truncatum* (Auyong, 2015). The identification of PuCut1 in *P. ulei* may be associated with its host specificity to *Hevea spp* and could explain why *P. ulei* is unable to infect *Hevea* leaves in advanced stages of phenological development.

Phytopathogenic fungi employ superoxide dismutases to protect themselves from reactive oxygen species (ROS) produced by the host and by their own metabolism (Warris & Ballou, 2019). ROS production generally occurs at the cytoplasmic level, however, a recent study described a superoxide dismutase is secreted to the extracellular space, where it degrades harmful fungal substances released in the host apoplast (Tian et al., 2021). We propose that the apoplastic superoxide

dismutase [Mn] protein found in *P. ulei* (PsSOD-Mn) could be critical in the detoxification of ROS, released by *H. brasiliensis*. Studies carried out by Koop et al., (2016) identified the overexpression of genes associated with the ROS production (HbRBOHA, HbRBOHB, HbRBOHC, HbRBOHD) in *H. brasiliensis* clones, which were either completely or partially resistant to *P. ulei*, compared to a susceptible clone. The expression changes were given 48 hours after being inoculated with a *P. ulei* strain.

Two effectors apoplastic putative of *P. ulei* were predicted as CAZymes. One of them is an acetyl xylan esterase corresponding to the Carbohydrate Esterase family (CE), and the other is a glycosyl hydrolase belonging to the family GH61. These CAZymes are associated with the host cell wall degradation in various filamentous fungal species (Vogel, 2008). Acetyl xylan esterase catalyzes an acetyl group's hydrolysis from xylan, the primary compound of hemicellulose, accounting for 20–30% of the cell wall of eudicotyledones and 50% in monocotyledones (Li et al., 2020). Glycoside hydrolase family 61 enhances the cellulose decomposition by direct oxidation (Žifčáková & Baldrian, 2012). As mentioned earlier, *P. ulei* penetrates the *H. brasiliensis* leaflets through the appressorium formation. However, it has also been described that at lower rates, *P. ulei* can also invade the host intercellular spaces without appressorium formation (Hashim et al., 1978). Whether or not the appressorium formation mediates the host colonization, CAZymes activity is essential for mediating host cell wall softness. One of our more interesting findings is the reduced number of CAzymes and effectors candidates with cellulase and xylanase activity that compounds the *P. ulei* secretome. Our results could explain both, the low percentage of *P. ulei* colonization without appressorium formation and the restriction of *P. ulei* to only infect leaves in early phonological development stages (A, B, C) (Guyot & Le Guen, 2017).

The *P. ulei* candidate effectors with similarity to *P. fulva* effectors (Ecp-5, Ecp-20-2, and Ecp-22), are part of a set of 70 effectors identified by transcriptomic and proteomic methods in the *Solanum lycopersicum* apoplast after to be inoculated with *P. fulva* strains (Mesarich et al., 2017). Ecp-5 is an effector that corresponds

with a conserved protein 115 aa, comprising six cysteineresidues whose deletion decreases the *P. fulva* virulence (Stergiopoulos & Wit, 2009). This effector has been associated with hypersensitive response in *Solanum pimpinellifolium* lineages, during the early stages of the *P. fulva* infection (Mesarich et al., 2017). The predicted effector Ecp-5 (termed TDFMu1) corresponds to a 129 aa protein that contains six cysteine residues. TDFMu1 was the first pathogenicity protein identified in the interaction of *H. brasiliensis* (clones IAN 710 and RRIM600) and *P. ulei* through cDNA-AFLPS markers and transcriptomic assays (Hurtado et al., 2015). The TDFMu1 expression was subsequently assessed by qRT-PCR, confirming the TDFMu1 gene upregulation in all the assessed clones i,e., susceptible *H. brasiliensis* clone, the partial resistance, and the total resistance, at 24 and 48 hpi, maintaining the expression rates until 144 hpi in total resistant clones (García, 2012). Through assembling the *P. ulei* genome, we have obtained the complete TDFMu1 gene sequence. This information is valuable for further assays to determine its role in the pathogenicity of *P. ulei* and to identify the corresponding resistance genes in *H. brasiliensis*.

Finally, a *P. ulei* effector candidate presented an identity percentage of 66.6% with Avr4 effector in *P. musae*. Avr4 effector is one of the few apoplastic effectors with a well-described and validated function, described in several Mycosphaerellaceae species (Zaccaron & Stergiopoulos, 2020; Chang et al., 2016; Esse et al., 2007). This effector has a length of 86 aa, and contains a CMB_14 domain that recognizes and kidnapped chitin molecules, preventing *L. solanum* defense response activation (Esse et al., 2007). In incompatible interactions, the Avr4 effector is recognized by Cf-4, an RLT type resistance gene triggering HR. Contrary, in compatible interaction, Avr4 prevents the release of chitin monomers, avoiding the recognition and the subsequent activation of the plant basal defense (PTI). Therefore, we suggest that PuAvr4 could play an essential role in the *P. ulei* virulence by promoting the infection and colonization of susceptible *Hevea spp*.

Table 1- 16. Apoplastic effectors candidates predicted in *P. ulei* genome.

N°	ID	Best blast homolog	Description	E-value	Size (aa)
1	jg3135.t1	KAF7190021.1	Acetylxylan esterase 2 [Pseudocercospora fuligena]	8,00E-110	228
2	jg5152.t1	KAF7192384.1	Cutinase 3 [Pseudocercospora fuligena]	5,00E-118	216
3	jg11381.t1	KX943057.1	Extracellular protein 20-2 (Ecp20-2) gene, complete cds [Passalora fulva strain 0WU]	2,00E-31	148
4	jg4921.t1	ABM47073.1	Extracellular protein 5 (Ecp-5) [Passalora fulva]	2,00E-12	129
5	jg3044.t1	XP_007922508.1	Glycoside hydrolase family 61 protein [Pseudocercospora fijiensis CIRAD86]	4,00E-141	258
6	jg1312.t1	KXT07835.1	Select seq gb Avr4 [Pseudocercospora musae]	5,00E-34	117
7	jg5675.t1	KZM22236.1	Hypothetical protein ST47_g6603 [Ascochyta rabiei]	5,00E-68	376
8	jg8564.t1	KXS96160.1	Hypothetical protein AC578_2665 [Pseudocercospora eumusae]	3,00E-40	114
9	jg1601.t1	KXT03241.1	Hypothetical protein AC578_4820 [Pseudocercospora eumusae]	2,00E-68	122
10	jg4508.t1	KXS95744.1	Hypothetical protein AC578_5300 [Pseudocercospora eumusae]	2,00E-73	193
11	jg8272.t1	KXT07360.1	Hypothetical protein AC578_538 [Pseudocercospora eumusae]	4,00E-47	162
12	jg8248.t1	KXT02753.1	Hypothetical protein AC578_5446 [Pseudocercospora eumusae]	5,00E-163	280
13	jg8960.t1	KXT05687.1	Hypothetical protein AC578_5577 [Pseudocercospora eumusae]	1,00E-52	116
14	jg1307.t1	KXS96648.1	Hypothetical protein AC578_6920 [Pseudocercospora eumusae]	2,00E-71	151
15	jg2297.t1	KXS94762.1	Hypothetical protein AC578_7625 [Pseudocercospora eumusae]	1,00E-135	265
16	jg1651.t1	KXT07842.1	Hypothetical protein AC579_1881 [Pseudocercospora musae]	5,00E-04	152

17	jg223.t1	KXS93493.1	Hypothetical protein AC579_2224 [Pseudocercospora musae]	8,00E-72	207
18	jg4263.t1	KXT10203.1	Hypothetical protein AC579_438 [Pseudocercospora musae]	6,00E-27	139
19	jg9575.t1	XP_664404.1	Hypothetical protein AN6800.2 [Aspergillus nidulans FGSC A4]	2,00E-40	198
20	jg11911.t1	OQO04011.1	Hypothetical protein B0A48_10654 [Rachicladosporium antarcticum]	2,00E-07	126
21	jg5595.t1	XP_021884475.1	Hypothetical protein BCR41DRAFT_347801 [Lobosporangium transversale]	7,00E-09	102
22	jg3771.t1	XP_023459240.1	Hypothetical protein CB0940_00438 [Cercospora beticola]	0.0	354
23	jg5423.t1	XP_023451268.1	Hypothetical protein CB0940_09931 [Cercospora beticola]	0.001	146
24	jg6769.t1	XP_023448439.1	Hypothetical protein CB0940_12139 [Cercospora beticola]	3,00E-34	141
25	jg8318.t1	XP_023448439.1	Hypothetical protein CB0940_12139 [Cercospora beticola]	3,00E-35	139
26	jg4440.t1	XP_023448439.1	Hypothetical protein CB0940_12139 [Cercospora beticola]	4,00E-35	143
27	jg11693.t1	XP_023448439.1	Hypothetical protein CB0940_12139 [Cercospora beticola]	9,00E-34	140
28	jg12255.t1	XP_023448439.1	Hypothetical protein CB0940_12139 [Cercospora beticola]	4,00E-19	148
29	jg12257.t1	XP_023448439.1	Hypothetical protein CB0940_12139 [Cercospora beticola]	1,00E-26	294
30	jg8748.t1	XP_023448439.1	Hypothetical protein CB0940_12139 [Cercospora beticola]	1,00E-23	142
31	jg12256.t1	XP_023448439.1	Hypothetical protein CB0940_12139 [Cercospora beticola]	9,00E-20	151
32	jg8889.t1	PPJ58006.1	Hypothetical protein CBER1_11801 [Cercospora berteroae]	4,00E-90	223
33	jg4746.t1	KAF7197604.1	Hypothetical protein HII31_01107 [Pseudocercospora fuligena]	7,00E-37	146

34	jg1570.t1	KAF7197604.1	Hypothetical protein HII31_01107 [Pseudocercospora fuligena]	8,00E-40	148
35	jg3597.t1	KAF7197248.1	Hypothetical protein HII31_01403 [Pseudocercospora fuligena]	1,00E-81	177
36	jg12469.t1	KAF7193647.1	Hypothetical protein HII31_04993 [Pseudocercospora fuligena]	2,00E-30	115
37	jg4739.t1	KAF7192376.1	Hypothetical protein HII31_06408 [Pseudocercospora fuligena]	3,00E-92	195
38	jg5039.t1	KAF7189516.1	Hypothetical protein HII31_09156 [Pseudocercospora fuligena]	5,00E-12	141
39	jg1575.t1	KAF7188888.1	Hypothetical protein HII31_09811 [Pseudocercospora fuligena]	4,00E-09	204
40	jg6430.t1	KAF7188652.1	Hypothetical protein HII31_09904 [Pseudocercospora fuligena]	2,00E-116	288
41	jg4189.t1	KAF7186110.1	Hypothetical protein HII31_12571 [Pseudocercospora fuligena]	4,00E-37	128
42	jg10087.t1	KAF7197209.1	Superoxide dismutase [Mn] [Pseudocercospora fuligena]	3,00E-145	238
43	jg5669.t1	XP_007924831.1	Uncharacterized protein MYCFIDRAFT_195328 [Pseudocercospora fijiensis CIRAD86]	1,00E-35	94
44	jg3884.t1	XP_007928545.1	Uncharacterized protein MYCFIDRAFT_88023 [Pseudocercospora fijiensis CIRAD86]	2,00E-86	179
45	jg11363.t1	XP_016756192.1	Uncharacterized protein SEPMUDRAFT_121575 [Sphaerulina musiva SO2202]	5,00E-05	135
46	jg2661.t1	XP_016756192.1	Uncharacterized protein SEPMUDRAFT_121575 [Sphaerulina musiva SO2202]	5,00E-05	136
47	jg6272.t1	_	Uncharacterized protein SEPMUDRAFT_121575 [Sphaerulina musiva SO2202]	_	139
48	jg9971.t1	_	without assignments	_	115
49	jg11361.t1	_	without assignments	_	133
50	jg2699.t1	_	without assignments	_	133
51	jg12481.t1	_	without assignments	_	125

52	jg3974.t1	_	without assignments	_	90
53	jg4197.t1	_	without assignments	_	190
54	jg4196.t1	_	without assignments	_	100

3.4. *P. ulei* transposable elements annotation

The TEs specific annotation results from the *P. ulei* genome allowed to classify the unclassified TEs identified during the genome masking. The TEs family curated consensus contributed in the classification of a total of 60’851 TEs sequences, which were clustered into 9 superfamilies (Figure 3-8). Among the TEs successfully classified, 51% (n = 31’088) were assigned to Class I and 48% (n= 29’599) to Class II. The most represented superfamilies for Classes I and II were *Gypsy* LTRs (RLG) with 22’844 assignments, and MITEs with 29’421, respectively (Figure 3-8).

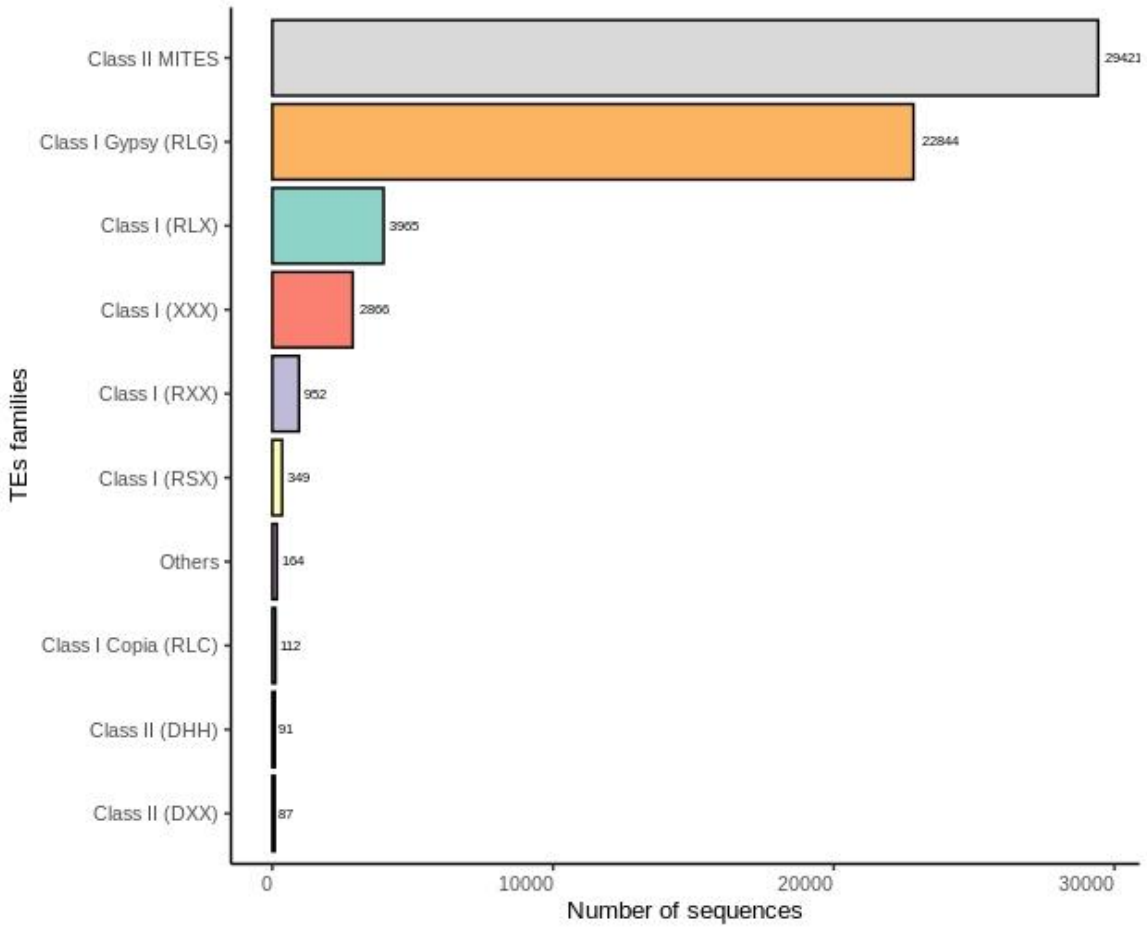


Figure 3-8. TEs superfamilies distribution in *P. ulei* genome

The *P. ulei* TEs classified into *Gypsy* (RLG) superfamily were mainly represented by RLG_Mira retrotransposons with 61% (13'990), RLG_Ginan with 29% (n = 6'685). The remaining, the TEs Class I, were assigned to the RLX, RSX, and unclassified XXX superfamilies being overrepresented by Alya, Chara and Ran retrotransposons, respectively. (Figure 3-9).

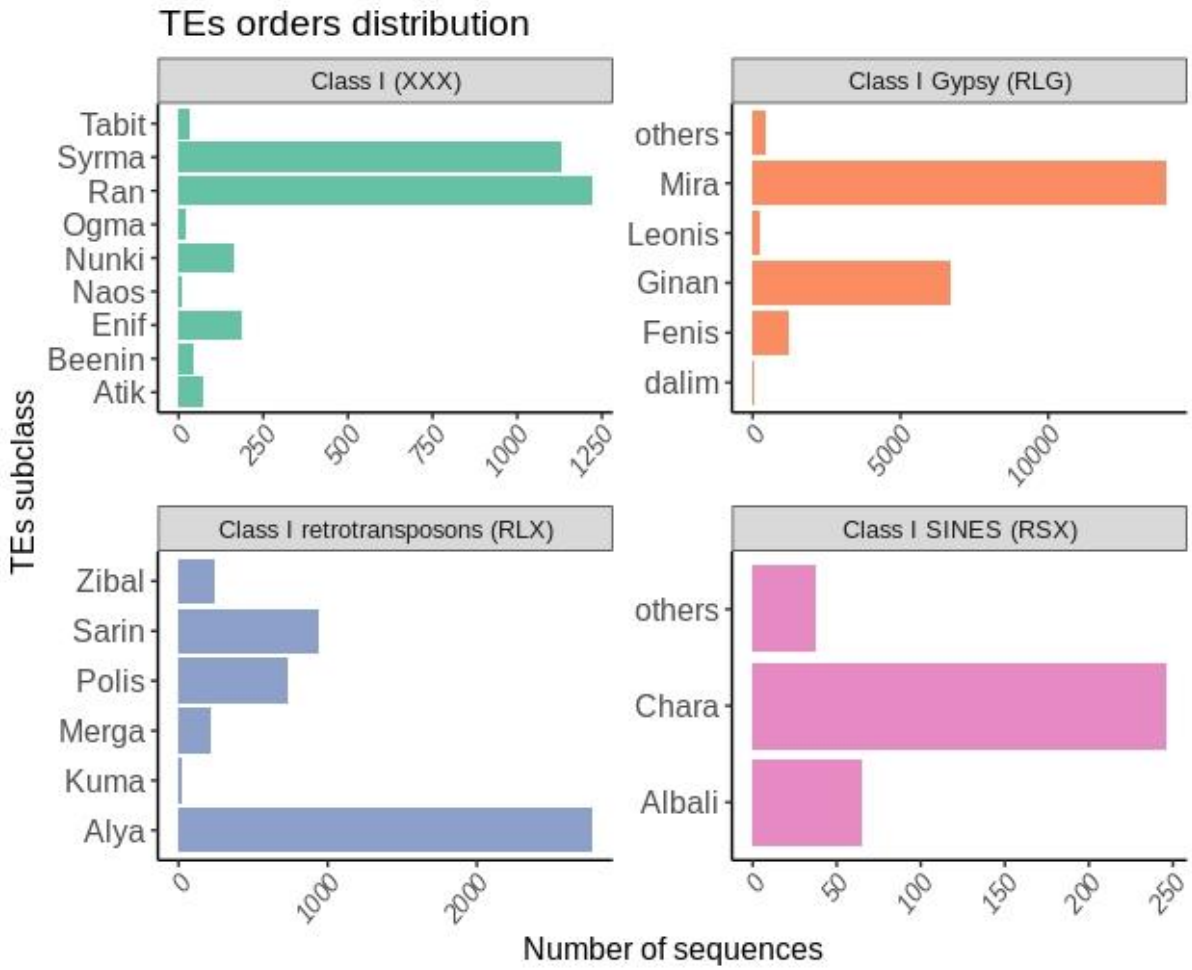


Figure 3-9. TEs orders distribution in *P. ulei* genome

The MITES (Miniature Inverted-repeat Transposable Elements) are class II non-autonomous elements that are often present in high numbers in the host genome of plants, animals and fungi (Yeadon & Catcheside, 1999; Feschotte & Mouche, 2000; Zhang, Arbuckle & Wessler, 2000; Feschotte, Swamy & Wessler, 2003). MITEs derive from diverse autonomous DNA transposons of the order of TIRs (terminal inverted

repeats) but have lost the coding region for transposase. Besides, MITEs structural characteristics like small size and its high number of copies have been described as facilitators for its retention and accumulation in the host genome compared with larger and low copy number elements which could be easily removed by drift or selection (Jiang et al., 2004). These results allowed us to infer that *P. ulei* genome shaping and adaptation against biotic and abiotic stress could be modeled by the presence of MITEs transposons.

Remarkably, the more abundant PFAM and GO categories identified in *P. ulei* genome functional annotation, ie., RNase H-like domains found in reverse transcriptase, reverse transcriptase (RNA-dependent DNA polymerase), and integrase zinc binding domain, might indicate that *P. ulei* TEs elements could be active. However, further studies that confirm the MITES activity and its effects in the *P. ulei* genome shape and adaptation are still needed.

3.5. Final remarks

In general, the *P. ulei* secretome was remarkably small with 756 compared with the sequenced *Pseudocercospora* species. The number of CAZymes in this genome is among the lowest in a set of 69 mycelial ascomycetes (Hane et al., 2020), and in the seven *Pseudocercospora* spp. sequenced genomes. The CAZymes annotation analysis allowed to stablish that the *P. ulei* lifestyle is monomerthroph, which is analogous to the biotrophic lifestyle (Hane et al., 2020). Besides, the CAZymes distribution revealed that the *P. ulei* carbohydrate degrading activity is focused on mild cellulose degradation. This could be compatible with the invasion mechanism that mediated by appressoria formation (Mendgen & Deiding, 1993).

*P. ule*i nutritional metabolism was also supported by the low number of gene clusters for secondary metabolites with fourteen gene clusters. Among them, we identified the presence of a melanin gene clusters supporting previous studies that describes the *P. ulei* invasion process is likely initiated by appressorium formation. Notably, genes cluster assigned to enzymes required for siderophore biosynthesis were absent in *P. ulei* genome. This finding might suggest that iron uptakes could be mediated by an alternative

system than siderophore-assisted iron uptake like RIA. Our functional annotation results support this hypothesis, since iron ion binding was one of the most abundant GO function category annotated in *P. ulei*.

Regarding the effector candidates, the secretome prediction pipeline allowed to identify a total of 113 putative effectors with a length between 100 and 358 amino acids. From them, 54 presented potential activity in the host apoplast. Among, the effector candidates without identified location, we found eight proteins assigned to transport functions such as exocytosis process. Exocytosis have been associated with fungal morphogenesis and pathogenicity processes (Chen et al. , 2015), and one diene lactone hydrolase protein related with hydrolyzation of aromatic-type compounds, that could be implied in the degradation of the scopoletin produced by the host during the pathogens colonization (Garcia, et al. 1995; Churngchow & Rattarasarn, 2001; Silva et al., 2002). Additionally, we found seven apoplastic effectors candidates with functional role in the host invasion process (Auyong, 2015), and the defense plant response inhibition (Chang et al., 2016). Of those we identified a cutinase3, a superoxide dismutase [Mn], two CAZymes comprehended by an acetyl xylan esterase 2, a glycosyl hydrolase of family 61, and four proteins with similarity to Avr4, Ecp5, Ecp-20-2, and Ecp-22 effectors.

Finally, the specific annotation of TEs in the *P. ulei* genome led to the identification of a significant proportion of unclassified TEs during the repetitive elements masking. Among the TEs successfully classified, 51% (n = 31’088) were assigned to Class I and 48% (n= 29’599) to Class II. The most represented superfamilies for Classes I and II were *Gypsy* LTRs (RLG) with 22’844 assignments, and MITEs with 29’421, respectively. These results reveled that *P. ulei* genome had a massive MITEs burst, which probably could be related with it genome oversize. Genomes sizes differ among Dothideomycetes with a strong correlation to transposable element content (Haridas et al., 2020). *Pseudocercospora* species show similar correlations, with the Sigatoka complex of banana pathogens carrying the biggest genomes. *P. fijiensis* and *P. musae* have 52.7 and 50.1% of TE content, respectively.

4. Conclusions and recommendations

4.1. Conclusions

Nowadays, assembly of a reference genome sequence for a species is critical for most basic and applied research, including the systematic identification of pathogenicity genes. In this study, we sequenced, assembled and annotated the genome of the poor studied fungus *P. ulei*, a limitant pathogen of natural rubber tree (*Hevea brasiliensis*). In this thesis, we present the first *P. ulei* high-quality genome assembly, the largest among Mycosphaerellaceae, with 93.8 Mbps, comprising 215 scaffolds, an N50 of 2.8 Mbps and a BUSCO gene completeness of 97.5%. We identified 12'745 protein-coding gene models in the *P. ulei* genome with 756 genes encoding secreted proteins and 113 genes encoding effector candidates. Most of the genome (80%) is composed of transposable elements, dominated by the Gypsy and MITEs superfamilies.

The results generated in this thesis supported that the *P. ulei* genome architecture is compatible with a one speed genome. The evaluation of the local density of TEs and genes throughout the *P. ulei* genome revealed that TEs are distributed in a scattered way presenting a decrease in the gene-rich regions (Figure 4-1). *P. ulei* TEs content and distribution is comparable with the highly compact genome of *Blumeria graminis forma specialis hordei,* a obligate biotrophic pathogen of barley crops that present 90% of the transposable elements (Spanu et al.,2010). The *Blumeria graminis* genome organization (retrotransposon proliferation, genome-size expansion, and gene losses) have been strongly correlated with its biotrophic lifestyle and species-specific adaptation events (Frantzeskakis et al., 2018). Similarly, the genome of *Sitophilus oryzae*, a weevil that affects the rice crop, presents a highly compact architecture. Interestingly, repetitive elements in the *S. oryzae* genome harbor non-coding DNA regions, especially intronic sequences generating introns with larger length (Parisot et al,. 2021). This pattern is also observed in *P. ulei* genome that have mean introns length of 386 bps. This value is

almost three times than the exhibited by the *P. fijiensis* and *P. musae*, the species with larger TEs content genomes in this genus.

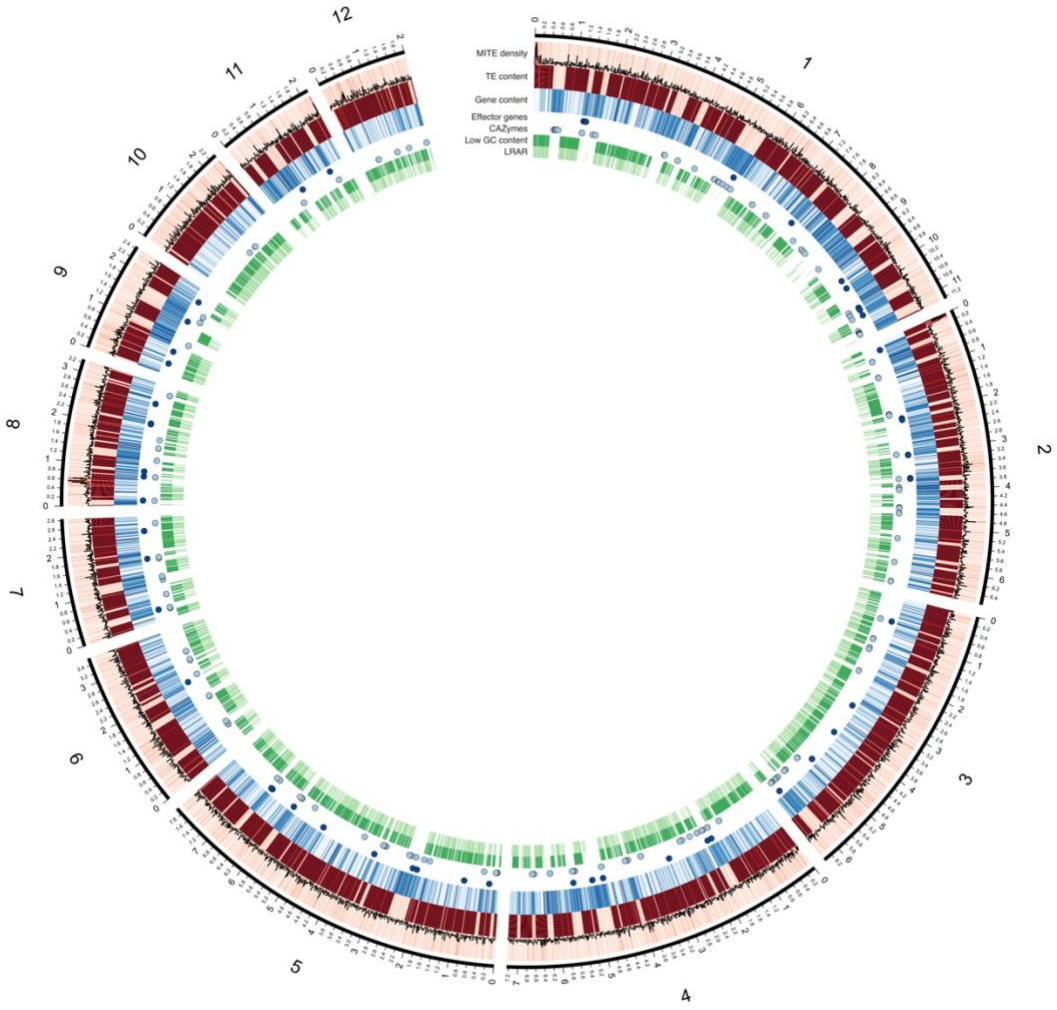


Figure 4-1: Circos plot of the largest scaffolds of *P. ulei* genome, illustrating the TEs distribution as well as other genome properties from outside to inside: 1. MITE copies histogram, 2. TE density (red), 3. gene density (blue), 4 position effectors (dark blue circle), 5. position CAZymes (light blue circle), 6. region with a GC lower than the mean GC content of the genome (dark green), 7. Large RIP affected regions (light green; according to TheRIPper).

The phylogenetic reconstruction enables to stablish that the closest relative of *P. ulei* is *P. camelliicola*, a pathogen of camellia cultivars. The topology based on the four loci markers showed that *P. ulei* was placed into a *Pseudocercospora* species group that has not been genetically explored yet, being the *P. ulei* genome, the first sequenced in this group. On the other hand, the phylogenomic analysis of the seven sequenced *Pseudocercospora* species, and 37 species outside *Pseudocercospora* genus allowed to define that *P. macadamia* is more taxonomically closed to the rubber pathogen than the black Sigatoka complex species. The genome of *P. ulei* contributes with new genome information in a group of *Pseudocercospora* that is still not fully explored from the genomics point of view, and could help answer evolutionary and biological unsolved questions for these plant pathogens.

The nutritional lifestyle of *P. ulei* was defined as a monomerthroph lifestyle, which is analogous to the biotrophic lifestyle (Hane et al.,2020). To date, *Passalora fulva* is the only species with biotrophic metabolism in the Mycosphaerellaceae family being one of the referent species to compare our pathogenicity findings. Additionally, both species *P. ulei* and *P. fulva* has an specialized metabolism, infecting only plant species within the same genus, i.e., *Hevea* and *Lycopersicum* respectively. On the other hand, the *Pseudocercospora* species included in this study have been classified as hemibiotrophic or necrotrophic, however, their genomes are the largest described within the Mycosphaerellaceae family being suitable models for genome architecture comparisons (Gonzalez et al., 2020; Wit et al., 2012).

One of the biggest motivation to develop this work, was having a signal about the molecular bases of *P. ulei* pathogenicity mechanisms and the interaction with its host *H. brasiliensis*. Therefore, we addressed the *P. ulei* functional annotation to identify genes that are required for the pathogenicity, and improve our understanding of the infection mechanisms of *P. ulei* on its host. In the next paragraphs we will conclude our main findings into a host-pathogen interaction model (Figure 4-2), using as a start point the preliminary findings of the natural

rubber group. The interaction model will define the possible molecular scenario given during the SALB disease development.

Previous efforts aimed to establish a model of the molecular relationship of *P. ulei*, and *H. brasiliensis* were performed by the natural rubber research group. The transcriptome analysis of the SALB interaction at different times after infection, enabled us to identify genes associated with the defense response of the plant, especially with the salicylic acid production. In this study, were identified a gene with homology to a polygalacturonase (Hb42704) and two genes with homology to glycosyl hydrolases from families 3 and 9 (Hb36621 and Hb44430) which were overexpressed in a partial resistant *H. brasiliensis* clone (Hurtado et al., 2015). From the pathogen perspective, this assay detected the expression of the aforementioned effector candidate TDFMu1. Based on this finding was proposed a second project aimed to study the apoplast proteome of the interaction between the susceptible *H. brasiliensis* clone (RRIM600) and with *P. ulei* at 9 post-infection days. 8 secreted proteins with sizes between 30 and 10 Kd were identified in the host apoplast. However, the determination of the complete sequence of them was not possible due to the lack of a reference genome of *P. ulei* (Mendez, 2017). Therefore, it was not feasible to propose a pathogenicity mechanism beyond the presumption of the fungal secretion of possible effectors to the apoplast host level.

The functional annotation of *P. ulei* genome predicted 756 proteins to be secreted, 113 classified as putative effectors including 54 presenting potential activity in the host apoplast. The number of effectors candidates was the lowest compared with the sequenced *Pseudocercospora* species (Gonzalez et al., 2021). Regarding the apoplastic effectors, these were lower than the identified in *P. fulva* which presented 70 candidates (Jaswal et al., 2020). The identification of a low number of effectors in *P. ulei*, could be due to the presence of effector proteins that are secreted by alternate secretion routes to those classically described, which could be excluded during the secretome prediction due to the absence of signal peptide (Jaswal et al. al., 2020).

The effector candidates of *P. ulei* for which was possible to assign function explain molecular events that take place during the penetration and host colonization. Taking into consideration our findings, we inferred that the SALB infection occurs from *P. ulei* viable conidia which is able to penetrate both susceptible and resistant *Hevea brasiliensis* plants in early phenological stages (A, B, C). The cuticle penetration happens through the formation of appressorium or not, in the second case the primary hyphae is able to penetrate the host cuticle (Figure 4-2, a).

Whether or not the appressoria formation, *P. ulei* secrete a limited number of CAZymes including glycosyl hydrolases of the family GH61 (jg3044.t1) (Figure 4-2, b). These enzymes are addressed to the cellulose degradation having the aim of softening the host cell wall during the invasion process, which is reinforced by the secretion of effectors such as cutinases (jg5152.t1), xyloglucan esterases (jg3135.t1). Once the cuticle and epidermis have been penetrated, the hyphae of the fungus grow through the intercellular spaces (Figure 4-2, c-d).

For a successful spreading into the host tissues, it is mandatory to counteract the plant immunity activation. For that end, Avr4-type effectors (jg1312.t1) are secreted to the apoplast with the aim of kidnapping chitin monomers or its derivates released during the invasion process Figure 4-2, e. The colonization process of *P. ulei* takes six days approximately and is defined by the presence of the signals and symptoms in the target tissues. In incompatible interaction, Avr4 is recognized by a host resistance protein present in the host membrane possibly of TLRs (Toll like Receptors) type (Rahman et al., 2013), and the host defense response is activated. This response is characterized by the release of CWDEs type GHs 3 and 9, and polygalacturinases (Hurtado et al., 2015) and by the ROS production (Koops et al., 2016) (Figure 4-2, e, f).

Additionally, we identified two *P. ulei* effectors addressed to minimize the host response activation harmful. One of them is an enzymes type oxidoreductases [Mn] (jg10087.t1) which is proposed to contribute in the ROS radical detoxification

(Figure 4-2, g) and the second is a dienolactone hydrolase (jg11150.t1) that is able to hydrolases aromatic compounds, we hypothesized that this hydrolase could has a role in the scopoletin degradation (Figure 4-2, g), a hidroxycumarine with antifungal activity produced by the host (Garcia, et al. 1995; Churngchow & Rattarasarn, 2001; Silva et al., 2002).

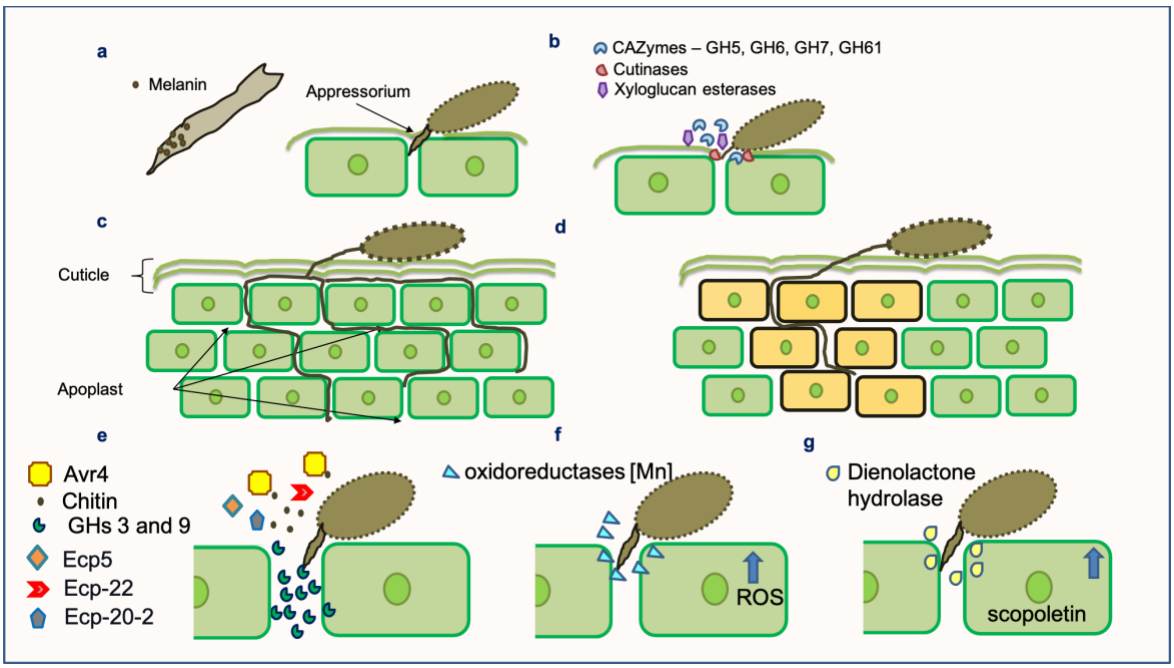


Figure 4-2: Hypothetical model of the molecular interaction given in SALB disease.

4.2. Future perspectives

1. We suggest sequencing new *P. ulei* isolates from other parts of the country or even from other countries such as Brazil and Guatemala. New sequenced *P. ulei* genomes would give rise to the population genomic studies contributing to assess the genome evolution and adaptation changes. Small and Large scale genome changes evaluated through macro y micro synteny analysis, could reveal the presence of chromosome structural variation, changes in the chromosome number and copy number variation events, as well as its correlation with the fungal fitness.
2. The transposable element's influence in *P. ulei* genome is definitively a subject that has to be reviewed with detail. It is quite important to determine if the MITES are active in *P. ulei* genome and also if they have any effect on the gene content modulation. Could be plausible that the low number of genes annotated in *P. ulei* genome was associated with TEs insertions events.
3. We recommended to perform functional validations for the Avr4 and Ecp5 apoplastic effector candidates confirming their function, sequence and organization. The identification of polymorphisms in the effector sequences could be crucial to contribute in the re-definition of *P. ulei* races .

A. Appendix 1: Accession numbers for the sequences of the four phylogenetic markers (actin, partial transcription elongation Factor 1-α (EF-1α) and ITS 1 and ITS2).

SPECIES	ITS	EF-1α	ACT
Pseudocercospora abelmoschi	GU269647	GU384365	GU320355
Pseudocercospora acericola	GU269650	GU384368	GU320358
Pseudocercospora ampelopsis	GU269830	GU384542	GU320534
Pseudocercospora angolensis	GU269836	GU384548	JQ325010
Pseudocercospora araliae	GU269652	GU384370	GU320360
Pseudocercospora arcuata	GU269850.1	JQ325006.1	GU320554.1
Pseudocercospora arecacearum	GU269655	GU384373	GU320363
Pseudocercospora assamensis	GU269656	GU384374	GU320364
Pseudocercospora atromarginalis	GU269658	GU384376	GU320365
Pseudocercospora balsaminae	GU269660	GU384379	GU320367
Pseudocercospora basiramifera	GU269661	GU384492	GU320368
Pseudocercospora basitruncata	GU269662	DQ211675	DQ147622
Pseudocercospora callicarpae	GU269663	GU384380	GU320369
Pseudocercospora catalpigena	GU269690	GU384406	GU320395
Pseudocercospora catappae	GU269669	GU384386	GU320375
Pseudocercospora cercidicola	GU269671	GU384388	GU320377
Pseudocercospora cercidis-chinensis	GU269670	GU384387	GU320376
Pseudocercospora cf. Kaki	GU269728	GU384441	GU320430
Pseudocercospora cheg	GU269673	GU384390	GU320379
Pseudocercospora chionanthi-retusi	GU269674	GU384391	GU320380
Pseudocercospora chrysanthemicola	GU269675	GU384392	GU320381
Pseudocercospora cladosporioides	GU269678	GU384395	GU320383
Pseudocercospora colombiensis	AY752149	DQ211660	DQ147639

Pseudocercospora contraria	GU269677	GU384394	GU320385
Pseudocercospora coprosmae	GU269680	GU384397	GU320386
Pseudocercospora cordiana	GU269681	GU384398	GU320387
Pseudocercospora coriariae	GU269682	GU384399	GU320388
Pseudocercospora cornicola	GU269683	GU384400	GU320389
Pseudocercospora corylopsidis	GU269721	GU384437	GU320425
Pseudocercospora cotoneastri	GU269685	GU384402	GU320391
Pseudocercospora crispans	GU269807	GU384518	GU320510
Pseudocercospora crocea	GU269792	GU384502	GU320493
Pseudocercospora crousii	GU269686	GU384403	GU320392
Pseudocercospora cruenta	GU269688	GU384404	JQ325012
Pseudocercospora cydoniae	GU269691	GU384407	GU320396
Pseudocercospora cymbidiicola	GU269692	GU384408	GU320397
Pseudocercospora davidiicola	GU269693	GU384409	GU320398
Pseudocercospora dendrobii	GU269696	GU384412	GU320401
Pseudocercospora destructiva	GU269694	GU384410	GU320399
Pseudocercospora dianellae	GU269695	GU384411	GU320400
Pseudocercospora dodonaeae	GU269697	GU384413	JQ325013
Pseudocercospora dovyalidis	GU269800	GU384513	GU320503
Pseudocercospora elaeocarpi	GU269701	GU384417	GU320405
Pseudocercospora eucalyptorum	GU269659	DQ211678	DQ147614
Pseudocercospora eupatoriella	GU269704	GU384420	GU320408
Pseudocercospora eustomatis	GU269705	GU384421	GU320409
Pseudocercospora exosporioides	GU269707	GU384423	GU320411
Pseudocercospora fijiensis (CIRAD 86)	KB446561	KB446555	KB446556
Pseudocercospora flavomarginata	GU269799	GU384512	DQ166513
Pseudocercospora fori	GU269806	GU384517	DQ147618
Pseudocercospora fraxinites	GU269672	GU384389	GU320378
Pseudocercospora fukuokaensis	GU269713	GU384429	GU320417
Pseudocercospora fuligena	GU269711	GU384427	GU320415
Pseudocercospora glauca	GU269715	GU384431	GU320419
Pseudocercospora gracilis	DQ267582	DQ211666	DQ147616
Pseudocercospora griseola f. griseola	GU269717	GU384433	GU320421
Pseudocercospora guianensis	GU269719	GU384435	GU320423
Pseudocercospora haiweiensis	GU269803	GU384514	GU320506
Pseudocercospora hakeae	GU269784	GU384495	JQ325017
Pseudocercospora humuli	GU269725	GU384439	GU320428
Pseudocercospora humulicola	GU269724	JQ324996	JQ325018
Pseudocercospora indonesiana	GU269735	GU384448	GU320437
Pseudocercospora ixorae	GU269726	GU384440	GU320429
Pseudocercospora jussiaeae	JQ324977	JQ324998	JQ325020
Pseudocercospora kakiicola	GU269729	GU384442	GU320431
Pseudocercospora kiggelariae	GU269730	GU384443	GU320432
Pseudocercospora latens	GU269732	GU384445	GU320434
Pseudocercospora leucadendri	GU269842	GU384555	GU320545
Pseudocercospora libertiae	GU269733	GU384446	GU320435

Pseudocercospora lilacis	GU269737	GU384449	GU320439
Pseudocercospora longispora	GU269734	GU384447	GU320436
Pseudocercospora lonicericola	GU269736	JQ324999	GU320438
Pseudocercospora luzardii	GU269738	GU384450	GU320440
Pseudocercospora lyoniae	GU269739	GU384451	GU320441
Pseudocercospora lythracearum	GU269740	GU384452	GU320442
Pseudocercospora lythri	GU269742	GU384454	GU320444
Pseudocercospora macrospora	GU269745	GU384457	GU320447
Pseudocercospora mali	GU269744	GU384456	GU320446
Pseudocercospora marginalis	GU269794	GU384504	GU320495
Pseudocercospora metrosideri	GU269746	GU384458	GU320448
Pseudocercospora musae	GU269747	GU384459	GU320449
Pseudocercospora myrticola	GU269749	GU384460	GU320451
Pseudocercospora nandinae	GU269750	GU384461	GU320452
Pseudocercospora natalensis	GU269780	JQ325000	DQ147620
Pseudocercospora nephrolepidis	GU269751	GU384462	GU320453
Pseudocercospora nogalesii	GU269752	GU384463	GU320454
Pseudocercospora norchiensis	GU269772	GU384484	GU320455.1
Pseudocercospora ocimi-basilici	GU269754	GU384465	GU320456
Pseudocercospora oenotherae	GU269856	GU384567	GU320559
Pseudocercospora palleobrunnea	GQ303288	GU384509	GU320500
Pseudocercospora pallida	GU269758	GU384469	GU320459
Pseudocercospora pancratii	GU269759	GU384470	GU320460
Pseudocercospora paraguayensis	DQ267602	DQ211680	DQ147606
Pseudocercospora pini-densiflorae	GU269760	GU384471	GU320461
Pseudocercospora plecthranthi	GU269791	GU384501	GU320492
Pseudocercospora pouzolziae	GU269761	GU384472	GU320462
Pseudocercospora profusa	GU269787	GU384497	GU320488
Pseudocercospora proteae	GU269808	GU384519	GU320511
Pseudocercospora prunicula	GU269676	GU384393	GU320382
Pseudocercospora pseudostigmina-platani	GU269857	GU384568	GU320560
Pseudocercospora puderi	GU269764	GU384476	GU320467
Pseudocercospora punctata	GU269765	GU384477	GU320468
Pseudocercospora purpurea	GU269783	GU384494	GU320486
Pseudocercospora pyracanthae	GU269767	GU384479	GU320470
Pseudocercospora ranjita	GU269790	GU384500	GU320491
Pseudocercospora ravenalicola	GU269810	GU384521	GU320513
Pseudocercospora rhabdothamni	GU269768	GU384480	GU320471
Pseudocercospora rhamnellae	GU269795	GU384505	GU320496
Pseudocercospora rhapisicola	GU269770	GU384482	GU320473
Pseudocercospora rhoina	GU269771	GU384483	GU320474
Pseudocercospora robusta	MH862873	DQ211683	DQ147617

Pseudocercospora rubi	GU269773	GU384485	GU320476
Pseudocercospora rumohrae	GU269774	GU384486	GU320477
Pseudocercospora sambucigena	GU269788	GU384498	GU320489
Pseudocercospora securinegae	GU269776	GU384487	GU320479
Pseudocercospora sordida	GU269777	GU384488	GU320480
Pseudocercospora stahlii	GU269813	GU384525	GU320515
Pseudocercospora stephanandrae	GU269814	GU384526	GU320516
Pseudocercospora subsessilis	GU269815	GU384527	GU320517
Pseudocercospora subtorulosa	GU269816	GU384528	GU320518
Pseudocercospora subulata	DQ303090	JQ325004	GU320519
Pseudocercospora tereticornis	JQ324982	GU384377	JQ325025
Pseudocercospora theae	GU269821	GU384534	GU320524
Pseudocercospora tibouchinigena	GU269822	GU384535	GU320525
Pseudocercospora timorensis	GU269823	GU384536	GU320526
Pseudocercospora variicolor	GU269826	GU384538	GU320530
Pseudocercospora viburnigena	GU269809	GU384520	GU320512
Pseudocercospora viticicola	GU269828	GU384540	GU320532
Pseudocercospora vitis	GU269829	GU384541	GU320533
Pseudocercospora weigelae	GU269831	GU384543	GU320535
Pseudocercospora xanthocercidis	JQ324983	JQ325005	JQ325026
Pseudocercospora xanthoxyli	GU269832	GU384544	GU320536
Pseudocercospora zelkovae	GU269833	GU384545	JQ325027
Pseudocercospora aeschynomenicola	KT290146	KT290200	KT313501
Pseudocercospora bixae	KT290153	KT290207	KT313508
Pseudocercospora boehmeriigena	KT290152	KT290206	KT313507
Pseudocercospora chamaecristae	KT290147	KT290201	KT313502
Pseudocercospora diplusodonii	KT290135	KT290189	KT313490
Pseudocercospora emmotunicola	KT290136	KT290190	KT313491
Pseudocercospora euphorbiacearum	KT313500	KT290199	KT313500
Pseudocercospora exilis	KT290139	KT290193	KT313494
Pseudocercospora manihotii	KT290144	KT290198	KT313499
Pseudocercospora perae	KT290132	KT290186	KT313487
Pseudocercospora Planaltinensis	KT290137	KT290191	KT313492
Pseudocercospora plumeriifolii	KT290138	KT290192	KT313493
Pseudocercospora plunkettii	KT290151	KT290205	KT313506
Pseudocercospora pothomorphes	KT290131	KT290185	KT313486
Pseudocercospora richardsoniicola	KT290154	KT290208	KT313509
Pseudocercospora Rigidae	KT290134	KT290188	KT313489
Pseudocercospora sennae-multijugae	KT290142	KT290196	KT313497

Pseudocercospora solani-pseudocapsicicola	KT290148	KT290202	KT313503
Pseudocercospora stizolobii	KT290143	KT290197	KT313498
Pseudocercospora struthanthi	KT290141	KT290195	KT313496
Pseudocercospora tecomicola	KT290156	KT290209	KT313511
Pseudocercospora wulffiae	KT290150	KT290204	KT313505
Pseudocercospora xylopiae	KT290133	KT290187	KT313488
Pseudocercospora eumusae	LFZN01000734.1	LFZN01000037.1	LFZN01000053.1
Pseudocercospora ulei	SCF 35	SCF 1	SCF 23
Pseudocercospora macadamia	WRNY01000139	KU878503	WRNY01000012
Pseudocercospora Piperis	JX875062	JX896123	-
Pseudocercospora prunicola	GU269676	GU384393	GU320382
Pseudocercospora pyracanthigena	GU269766	GU384478	GU320469
Pseudocercospora Sawadae	GU269775	-	GU320478
Pseudocercospora sp CBS 110998	GU269778	GU384489	GU320481
Pseudocercospora trinidadensis	KT290157	KT290210	-
Pseudocercospora udagawana	GU269824	GU384537	GU320527
Pseudocercospora vassobiae	KT290155	-	KT313510
Pseudocercospora ershadii	KM452865	KM452887	KM452842
Pseudocercospora fuliginosa	LC515776	LC515789	LC515779
Pseudocercospora kobayashiana	LC511998	LC515780	LC512004
Pseudocercospora diospyriphila	LC512003	LC515790	GU512009
Pseudocercospora paederiae	GU269757	GU384468	-
Pseudocercospora rhododendri-indici	GU269722	-	GU320426
Pseudocercospora snelliana	GU269731	GU384444	GU320433
Pseudocercospora thailandica	-	GU384533	GU320523
Pseudocercospora chibaensis	KX462584	KX462670	KX462551
Pseudocercospora cyathicola	JF951139	KX462673	KX462554
Pseudocercospora daphniphylli	KX462587	KX462674	KX462555
Pseudocercospora elaeocarpicola	KX462588	KX462675	KX462556
Pseudocercospora Eriobotryae	KX462589	KX462676	KX462557
Pseudocercospora eriobotryicola	KX462589	KX462677	KX462558
Pseudocercospora eupatorii-formosani	KX462591	KX462678	KX462559
Pseudocercospora hachijokibushii	KX462593	KX462680	KX462561
Pseudocercospora hiratsukana	KX462594	KX462681	KX462562
Pseudocercospora houttuyniae	KX462595	KX462682	KX462563
Pseudocercospora imazekii	KX462607	KX462693	KX462574
Pseudocercospora izuohshimense	KX462597	KX462684	KX462565
Pseudocercospora kadsurae	KX462598	KX462685	KX462566

Pseudocercospora madagascariensis	GQ852767	KF253265	KF253625
Pseudocercospora naitoi	KX462599	KX462686	KX462567
Pseudocercospora nephrolepidicola	HQ599590	KX462688	KX462569
Pseudocercospora neriicola	KJ869165	KJ869240	KJ869231
Pseudocercospora photiniae	KX462604	KX462690	KX462571
Pseudocercospora punicae	KX462606	KX462692	KX462573
Pseudocercospora tinea	KX462608	KX462696	GU320499
Pseudocercospora violamaculans	KX462610	KX462698	KX462579
Pseudocercospora xenosyzygiicola	KX462611	KX462699	KX462580
Pseudocercospora pyricola	KY048161.1	KY048164.1	KX462580.1
Pseudocercospora airliensis	KM055429.1	KM055436.1	-
Pseudocercospora jagerae	KM055431.1	KM055438.1	-
Pseudocercospora proiphydis	KM055430.1	KM055437.1	-
Pseudocercospora iwakiensis	KX462607.1	KX462693.1	KX462574.1
Pseudocercospora piricola	KY048161.1	KY048164.1	KY048162.1
Pseudocercospora aristoteliae	MK432806.1	-	-
Pseudocercospora amelanchieris	KX462583.1	KX462669.1	KX462550.1
Pseudocercospora athyrii	KJ201932.1	-	-
Pseudocercospora beijingensis	MH255813.1	MH255819.1	MH392526.1
Pseudocercospora bischofiae	KC677896.1	-	-
Pseudocercospora brackenicola	KT037524.1	KT037484.1	-
Pseudocercospora breonadiae	MH107913.1	MH108026.1	MH107985.1
Pseudocercospora buteae	KC677909.1	-	-
Pseudocercospora camelliicola	KJ201933.1	-	-
Pseudocercospora carbonacea	KC677897.1	-	-
Pseudocercospora casuarinae	HQ599603.1	-	-
Pseudocercospora chiangmaiensis	MH863288.1	KF903177.1	KF903544.1
Pseudocercospora christellae	KC677898.1	-	-
Pseudocercospora circumscissa	KX853044.1	KX853061.1	-
Pseudocercospora cladrastidis	AB694923.1	-	-
Pseudocercospora clematidis	DQ303072.1	JX901677.1	JX902133.1
Pseudocercospora clerodendri-hastati	LC146754.1	-	-
Pseudocercospora cyatheae	KJ201935.1	-	-
Pseudocercospora depazeoides	LT160035.1	-	-
Pseudocercospora dingleyae	KX287299.1	-	-
Pseudocercospora duabangae	KC677899.1	-	-
Pseudocercospora ebulicola	KJ201936.1	-	-
Pseudocercospora elaeocarpi	GU269701.1	GU384417.1	GU320405.1
Pseudocercospora elaeodendri	KC172073.1	-	-
Pseudocercospora euonymi	MH255812.1	MH255818.1	MH392525.1
Pseudocercospora fatouae	DQ303076.1	-	-

Pseudocercospora fici	KJ412456.1	-	-
Pseudocercospora ficisepticae	MW063148.1	-	-
Pseudocercospora ginkgoana	JX134048.1	-	-
Pseudocercospora gmelinae	KC677901.1	-	-
Pseudocercospora hamiltoniani	MN737835.1	-	-
Pseudocercospora heteropyxidicola	MN562151.1	-	-
Pseudocercospora hibbertiae-asperae	AF488743.1	-	-
Pseudocercospora holarrhenae	KC677902.1	-	-
Pseudocercospora jahnii	KM393283.1	-	-
Pseudocercospora kamalii	JF824126.1	-	-
Pseudocercospora leandrae-fragilis	KY574288.1	-	-
Pseudocercospora lindericola	KX853047.1	-	-
Pseudocercospora ygodiicola	KT037526.1	-	-
Pseudocercospora macarangae	KC677905.1	-	-
Pseudocercospora maetaengensis	MN648323.1	-	-
Pseudocercospora malloticola	KC677906.1	-	-
Pseudocercospora mapelanensis	KM203118.1	-	-
Pseudocercospora mazandaranensis	KM452855.1	KM452877.1	KM452832.1
Pseudocercospora microlepiae	KR348740.1	-	-
Pseudocercospora micromeli	KC677910.1	-	-
Pseudocercospora mombin	KC677907.1	-	-
Pseudocercospora nelumbonicola	KY304493.1	-	-
Pseudocercospora nodosa	MF951367.1	-	-
Pseudocercospora nymphaeacea	KT074354.1	KT074355.1	KT074353.1
Pseudocercospora oenotherae	MK863390.1	GU384466.1	GU320559.1
Pseudocercospora opuntiae	KF975410.1	-	-
Pseudocercospora paranaensis	KT037523.1	KT037483.1	KT037604.1
Pseudocercospora parapseudarthriae	KJ869151.1	-	-
Pseudocercospora pittospori	MK210511.1	-	-
Pseudocercospora platylobii	AY260089.1	-	-
Pseudocercospora protearum	AY251107.2	-	-
Pseudocercospora pseudomyrticola	MK876405.1	MK876499.1	MK876461.1
Pseudocercospora pteridicola	KR348738.1	-	-
Pseudocercospora pteridophytophila	KJ201938.1	-	-
Pseudocercospora puerariicola	MG601511.1	-	-

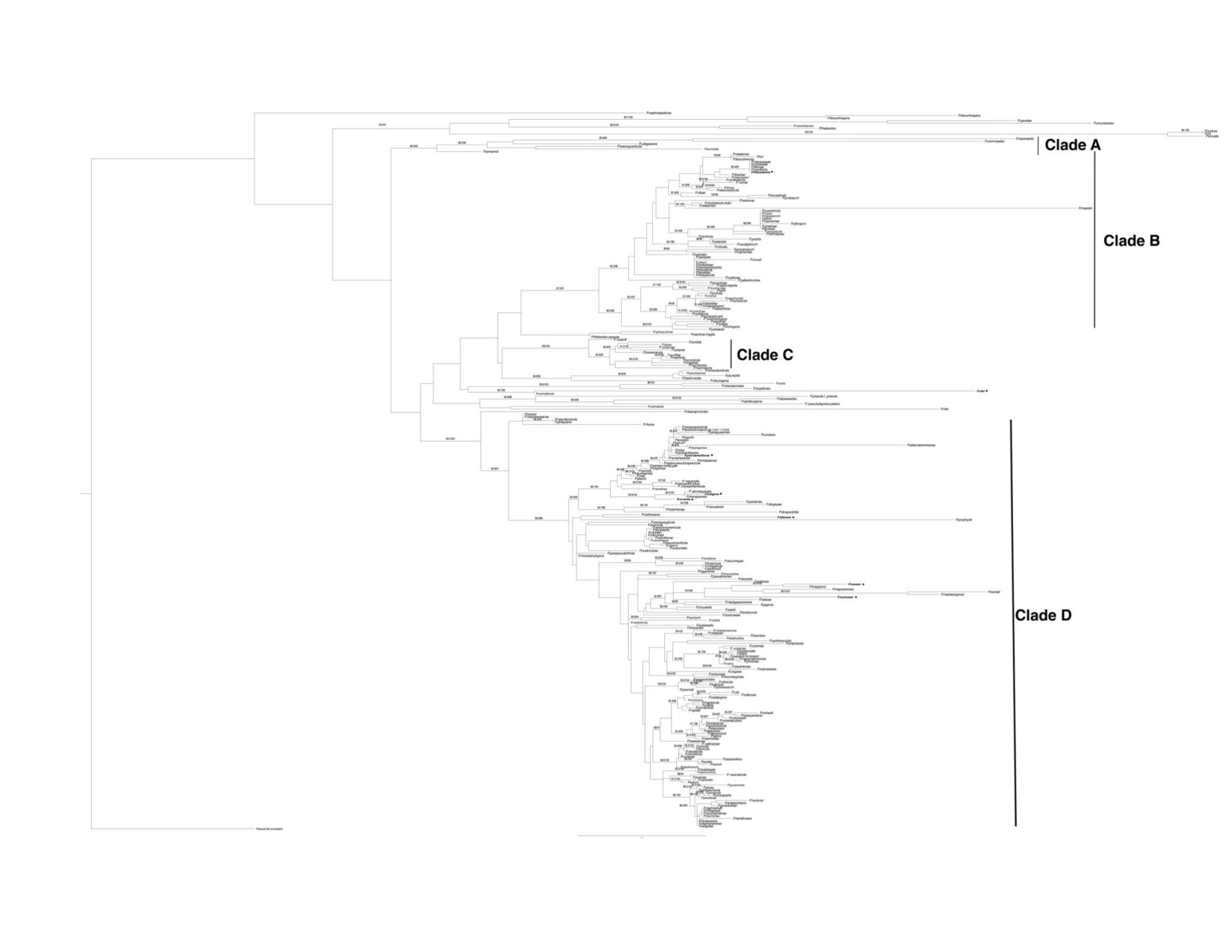
Pseudocercospora rhododendrigena	LC146756.1	-	-
Pseudocercospora rosae	MG828952.1	-	-
Pseudocercospora schizolobii	JQ676195.1	JX901695.1	KF253628.1
Pseudocercospora serpocaulonicola	KT037525.1	KT037485.1	KT037607.1
Pseudocercospora sphaerulinae	JX901791.1	KF903215.1	-
Pseudocercospora ubsynnematosa	KF686808.1	-	-
Pseudocercospora syzygiicola	AF309600.1	-	-
Pseudocercospora tabernaemontanae	KC677911.1	-	-
Pseudocercospora utamarindi	KP744461.1	-	-
Pseudocercospora thelypteridis	KT037521.1	-	-
Pseudocercospora tibouchinicola	KF686809.1	-	-
Pseudocercospora tibouchinigena	GU269822.1	-	-
Pseudocercospora trichogena	KT037520.1	KT037480.1	KT037601.1
Pseudocercospora wedeliae	KJ201940.1	-	-

B. Appendix 2: Accession numbers for the genome sequences of species within Mycosphaerellaceae used for phylogenomic analyses.

Species name	accesion number	Genome Size
Cercospora berteroae	Cbert_1	33,892,839
Cercospora beticola	CB0940_V2	35,260,078
Cercospora brassicicola	ASM1336524v1	38,324,398
Cercospora canescens	C_canescens_MTCC-10836_BHU_v01	33,967,224
Cercospora cf. flagellaris	ASM535688v1	33,240,740
Cercospora cf. sigesbeckiae	ASM221750v1	34,940,764
Cercospora citrullina	ASM1336519v1	32,813,481
Cercospora kikuchii	ASM535685v1	33,223,116
Cercospora nicotianae	CNIC01	33,372,350
Cercospora sesami	ASM1336523v1	34,336,664
Cercospora sojina	ASM429982v1	40,115,407
Cercospora sojina	ASM429982v1	40,115,976
Cercospora zeina	ASM284461v1	40,755,333
Dothistroma pini	ASM211635v1	29,984,598
Dothistroma septosporum	ECU13.long	30,196,509
Exutisphaerella laricina	ASM50438v2	26,520,711
Lecanosticta acicola	ASM50434v2	28,442,312
Microcyclosporella mali	ASM278598v1	27,977,477
Mycosphaerella arachidis	ASM129726v1	33,245,410
Mycosphaerella populi	ASM215340v1	30,013,927
Mycosphaerelloides madeirae	ASM278599v1	33,743,376

Nothophaeocryptopus gaeumannii	ASM211638v1	33,964,036
Pallidocercospora crystallina	ASM366608v1	36,952,682
Passalora fulva	CfCabog12	61,103,943
Passalora sequoiae	ASM1324884v1	31,768,716
Pseudocercospora cruenta	SM1336520v1	55,623,252
Pseudocercospora eumusae	ASM157823v1	47,119,461
Pseudocercospora fijiensis	Mycfi2	74,141,167
Pseudocercospora macadamiae	ASM1297840v1	40,070,143
Pseudocercospora musae	ASM157822v1	60,439,160
Pseudocercospora pini-densiflorae	ASM50436v2	43,513,371
Pseudocercospora ulei	This study	93,730,151
Ramularia coccinea	ASM415524v1	32,744,270
Ramularia collo-cygni	version 1	32,254,038
Ramularia endophylla	ASM211639v1	41,329,147
Sphaerulina musiva	Septoria musiva SO2202 v1.0	29,352,103
Sphaerulina populicola	sPop_v1	33,188,813
Zasmidium angulare	ASM278604v1	37,899,076
Zasmidium cellare	Zasce1	38,247,703
Zasmidium citrigriseum	ASM278602v1	44,999,914
Zymoseptoria ardabiliae	ASM22376v2	31,136,334
Zymoseptoria brevis	ASM96659v1	31,914,300
Zymoseptoria passerinii	ASM22382v2	28,786,283
Zymoseptoria pseudotritici	ASM22368v2	32,796,701
Zymoseptoria tritici	ASM293742v1	41,953,157

C. Appendix 3: Accession numbers for the genome sequences of species within Mycosphaerellaceae used for phylogenomic analyses


Clade A
Clade B
Clade C
Clade D

References

- Akinsanmi, O and Carvalhais L. (2020) Draft Genome of the Macadamia Husk Spot Pathogen, Pseudocercospora macadamiae. Phytopathology. 110:1503-1506 https://doi.org/10.1094/PHYTO-12-19-0460-A
- Albera S. (2016). Biotrophic Fungi Infection and Plant Defense Mechanism. Journal of Plant Pathology & Microbiology, 7(9), 1–6.
- Alic, A., Ruzafa, D., Dopazo, J., Blanquer, I. (2016). Objective review of de novo stand-alone error correction methods for NGS data. Wiley Interdisciplinary Reviews: Computational Molecular Science, 6(2), 111–146. doi:10.1002/wcms.1239
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410.
- Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc
- Angelo, S., Yamagishi, B., Cruz, C., Silva, G, & Gasparotto, L. (2020). Differential expression and structural polymorphism in rubber tree genes related to South American leaf blight resistance. Physiological and Molecular Plant Pathology, 110, 101477. doi:10.1016/j.pmpp.2020.101477
- Asai, S., & Shirasu, K. (2015). Plant cells under siege: plant immune system versus pathogen effectors. Current Opinion in Plant Biology, 28, 1–8. doi:10.1016/j.pbi.2015.08.008
- Auyong, S. (2015). The Role of Cutinase and its Impact on Pathogenicity of Colletotrichum truncatum. Journal of Plant Pathology & Microbiology, 06(03). https://doi.org/10.4172/2157-7471.1000259
- Bakhshi, Marzanlou, M., Babai-Ahari, A., Groenewald, J., Crous, P. (2014). Multi-gene analyses of Pseudocercospora spp. from Iran. Phytotaxa 184: 245 – 264
- Ballester, A., Marcet-Houben, M., Levin, E., Sela, N., Selma-Lázaro, C., Carmona, L., Wisniewski, M., Droby, S., González, L., Gabaldón, T. (2015). Genome, Transcriptome, and Functional Analyses of Penicillium expansum Provide New Insights Into Secondary Metabolism and Pathogenicity. Molecular Plant-Microbe Interactions, 28(3), 232–248. doi:10.1094/MPMI-09-14-0261-FI
- Bankevich, A., Nurk, S., Antipov, D., Gurevich, A., Dvorkin, M., Kulikov, A. S., Lesin, V., Nikolenko, S., Pham S., Prjibelski A., Pyshkin A., Sirotkin A., Vyahhi N., Tesler G., Alekseyev M. A., Pevzner, P. A. (2012). SPAdes: A New Genome Assembly

Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology

- Bao, W., Kojima, K., Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015;6(1):11. Available from: http://www.mobilednajournal.com/content/6/1/11.
- Bayry, A., Guijarro, J., Sunde, M., Latgé, J. (2012) Hydrophobins Unique Fungal In Biological Adhesives Vol. 107, pp. 25–55
- Beacham, G. M., Partlow, E. A., Lange, J. J., & Hollopeter, G. (2018). NECAPs are negative regulators of the AP2 clathrin adaptor complex. ELife, 7. https://doi.org/10.7554/eLife.32242
- Beckerman, L., and Ebbole, D. (1996). MPG1, a gene encoding a fungal hydrophobin of Magnaporthe grisea, is involved in surface recognition. Mol. Plant-Microbe Interact. 9:450-456.
- Bedre, R., Avila, C., Mandadi, K. (2021). HTSeqQC: A Flexible and One-Step Quality Control Software for High-throughput Sequence Data Analysis. BioRxiv https://doi.org/10.1101/2020.07.23.214536
- Benavidez. T, Jaron. K Schatz. Michael . (2020). GenomeScope 2.0 and Smudgeplot for reference- free profiling of polyploid genomes NATURE COMMUNICATIONS (2020)11:1432:https://doi.org/10.1038/s41467-020-14998-3
- Berlin, K., Koren, S., Chi, X., Drake, J., Landolin, J.,Phillippy, A (2015). Assembling large genomes with single-molecule sequencing and locality-sensitive hashing Nature Biotechnology 33, 623–630
- Bigeard, J., Colcombet, J., Heribert, H. (2015). Signaling Mechanisms in Pattern-Triggered Immunity (PTI). Molecular Plant 8, 521–539
- Blackwell, M. (2011). The Fungi: 1, 2, 3...5.1 million species? American Journal of Botany 98:426-438
- Blango, M., Kniemeyer, O., Brakhage, A., Sheppard, D. (2019). Conidial surface proteins at the interface of fungal infections. PLOS Pathogens, 15(9), e1007939–. doi:10.1371/journal.ppat.1007939
- Blatzer, M., Binder, U., Haas, H. (2011). The metalloreductase FreB is involved in adaptation of Aspergillus fumigatus to iron starvation. , 48(11), 0–1033. doi:10.1016/j.fgb.2011.07.009
- Blin, K., Medema, M., Kazempour, D., Fischbach, M., Breitling, R., Takano, E., et al. (2013). antiSMASH 2.0 a versatile platform for genome mining of secondary metabolite producers. Nucleic Acids Res. 41(1), Pages 204-212.
- Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014 Aug 1;30(15):2114-20. doi: 10.1093/bioinformatics/btu170.
- Boller, T., Yang, S. (2009). Innate immunity in plants: An arms race between pattern recognition receptors in plants and effectors in microbial pathogens. Science. May 8; 324(5928): 742–744.
- Bolton, M., van Esse, R., Vossen, J., de Jonge, R et al. (2008). The novel Cladosporium fulvum lysin motif effector Ecp6 is a virulence factor with orthologues in other fungal species. Molecular Microbiology 69(1), 119–136
- Breen, J., Wicker, T., Kong, X., Zhang, J., Ma, W., Paux, E., et al. (2010). A highly conserved gene island of three genes on chromosome 3B of hexaploid wheat: diverse gene function and genomic structure maintained in a tightly linked block. BMC Plant Biol. 2010;10:98. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20507561.

References

- Breen J, Wicker T, Kong X, Zhang J, Ma W, Paux E, et al. A highly conserved gene island of three genes on chromosome 3B of hexaploid wheat: diverse gene function and genomic structure maintained in a tightly linked block. BMC Plant Biol. 2010;10:98. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20507561
- Brown, A., Antoniw, J., & Hammond-Kosack, K. E. (2012). The Predicted Secretome of the Plant Pathogenic Fungus Fusarium graminearum: A Refined Comparative Analysis. PLoS ONE, 7(4), e33731. doi:10.1371/journal.pone.0033731
- Buermans, J., & Dunnen, T. (2014). Next generation sequencing technology: Advances and applications. BBA - Molecular Basis of Disease, 1842(10), 1932–1941.
- Bushnell, B. (2014). BBMap: A Fast, Accurate, Splice-Aware-aligner. LBNL Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, USA.
- Butler, E & Jones, S. (1949) Tomato Leaf Mould, Cladosporium fulvum Cooke. London: Macmillan.
- Camargo, A.P., Marin, F.R., Camargo, M.B.P., 2003. Zoneamento Climático da Hevea no Brasil, Documento 24. Embrapa, Campinas, SP, ISSN 0103-78110.
- Cantarel, L., Coutinho, M., Rancurel, C., Bernard, T., Lombard,V., Henrissat,B. (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. Nucleic Acids Res., 37, D233–D238.

- Capella, S., Silla, J., Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 25(15): Pages 1972-3.
- Casadevall, A., Rosas, L., Nosanchuk, D. (2000). Melanin and virulence in Cryptococcus neoformans. Curr Opin Microbiol. 2000 Aug;3(4):354-8. doi: 10.1016/s1369-5274(00)00103-x. PMID: 10972493.
- Casadevall, A. (2007).Determinants of virulence in the pathogenic fungi. Fungal Biol Rev. 2007 November ; 21(4): 130–132. doi:10.1016/j.fbr.2007.02.007.
- Castro., E, Sigrist, A., Gattiker, A., Bulliard, V., Langendijk-Genevaux, S., Gasteiger, E et al. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Res. 2006;34(Web Server issue):W362-W5. doi: 10.1093/nar/gkl124. PubMed PMID: PMC1538847.
- Cayla, V., Petch, D. (1911). La lutte contre les maladies cryptogamiques dans les plantations d'Hévéa Jour, Agro. Trop,. Ann. 11 p 329-335. Review of Petch, Physiology and Diseases of Hevea brasiliensis. London 1911
- Chang, T., Salvucci, A., Crous, P., Stergiopoulos, I. (2016) Comparative Genomics of the Sigatoka Disease Complex on Banana Suggests a Link between Parallel Evolutionary Changes in Pseudocercospora fijiensis and Pseudocercospora eumusae and Increased Virulence on the Banana Host. PLoS Genetic 12(8)
- Chee, K.H.; Holliday, P. (1986). Enfermedad suramericana de la hoja del hule (caucho) Hevea. Instituto para la Investigación y Desarrollo del Hule de Malasia, MRRDB. Monografía n° 13. Presentado en la serie técnica n° 37 Avances de la investigación en caucho natural. CONIF. 1997 Santa Fe de Bogotá.
- Chee, K. H. (1976). Assessing susceptibility of Hevea clones to *Microcyclus ulei*. Annals Of Applied Biology, 84, 135-145
- Chen, X., Ebbole, D., Wang, Z. (2015). The exocyst complex: delivery hub for morphogenesis and pathogenesis in filamentous fungi. Current Opinion in Plant Biology, 28(), 48–54. doi:10.1016/j.pbi.2015.09.003

- Chen, Y., Nie, F., Xie, SQ. et al. (2021). Efficient assembly of nanopore reads via highly accurate and intact error correction. Nat Commun 12, 60. https://doi.org/10.1038/s41467-020-20236-7
- Chikhi, R., & Medvedev, P. (2014) Informed and automated k-mer size selection for genome assembly, Bioinformatics, Volume 30, Issue 1, 1 January 2014, Pages 31–37
- Chin, et al. (2013). Non Hybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nature Methods. 10(6), 563.
- Chinchilla, D., Bauer, Z., Regenass, M., Boller, T. & Felix, G. (2006). The Arabidopsis Receptor Kinase FLS2 Binds flg22 and Determines the Specificity of Flagellin Perception. The Plant Cell Online, 18, 465-476.
- Chin et al. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. Nature Methods. 13(12), 1050.
- Chisholm, T., Coaker, G., Day, B., Staskawicz, B. (2006). Host–microbe interactions: shaping the evolution of the plant immune response. Cell 124: 803–814.
- Choi, J., & Kim, S. (2017). A genome Tree of Life for the Fungi kingdom. Proceedings of the National Academy of Sciences, (), 201711939–. doi:10.1073/pnas.1711939114
- Churngchow, N., Rattarasarn, M. (2001). Biosynthesis of scopoletin in Hevea brasiliensis leaves inoculated with Phytophthora palmivora. , 158(7), 0–882. doi:10.1078/0176-1617-00230
- Collemare, J., Griffiths, S., Iida, Y., Karimi, M., Battaglia, E., Cox, R., & de Wit, P. (2014). Secondary Metabolism and Biotrophic Lifestyle in the Tomato Pathogen Cladosporium fulvum. PLoS ONE, 9(1), e85877. doi:10.1371/journal.pone.0085877
- Compagnon P (1986) In: Maisonneuve GP, Larose (eds) Le caoutchouc naturel. P3
- Compeau, E., Pevzner, A., & Tesler, G. (2011). How to apply de Bruijn graphs to genome assembly. Nature biotechnology, 29(11), 987–991. https://doi.org/10.1038/nbt.2023
- Cornish K, Siler DJ, Grosjean OK, Godman N (1993). Fundamental similarities in rubber particle architecture and function in three evolutionarily divergent plant species. J. Nat. Rubber Res. 8:275-285.
- Costa, R., Resende, M., Araujo, AJ., Gonçalves, P., Higa, A. (2000). Selection and genetic gain in rubber tree (Hevea) populations using a mixed mating system. Genet. Mol. Biol. 23(3): 671-679.
- Couturier M, Navarro D, Olive C, Chevret D, Haon M, Favel A, Lesage-Meessen L, Henrissat B, Coutinho PM, Berrin JG: Post-genomic analyses of fungal lignocellulosic biomass degradation reveal the unexpected potential of the plant pathogen Ustilago maydis. BMC Genomics 2012, 13:57.
- Crescente, M., Zavallo, D., Helguera, M., Vanzetti, S. (2018). MITE Tracker: an accurate approach to identify miniature inverted-repeat transposable elements in large genomes. BMC Bioinformatics. 2018;19(1):348. Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2376-y
- Croll, D., & McDonald, B. (2012) The Accessory Genome as a Cradle for Adaptive Evolution in Pathogens. PLoS Pathog 8(4): e1002608. doi:10.1371/ journal.ppat.1002608
- Cross, A.S. (2008). What is a virulence factor? Critical care 12(6): 196.
- Crous, W., Braun, U., Hunter, GC., et al. (2013). Phylogenetic lineages in Pseudocercospora. Studies in Mycology 75: 37–114.

References

- Daboussi, J., & Capy, P. (2003). Transposable Elements in Filamentous Fungi. Annual Review of Microbiology, 57(1), 275–299. doi:10.1146/annurev.micro.57.030502.091029
- Damien J. Fleetwood, Anar K. Khan, Richard D. Johnson, Carolyn A. Young, Shipra Mittal, Ruth E. Wrenn, Uljana Hesse, Simon J. Foster, Christopher L. Schardl, Barry Scott, Abundant Degenerate Miniature Inverted-Repeat Transposable Elements in Genomes of Epichloid Fungal Endophytes of Grasses, Genome Biology and Evolution, Volume 3, 2011, Pages 1253–1264,
- De Jonge, R., Van Esse, P., Kombrink, A., Shinya, T., Desaki, Y., Bours, R., Van der Krol, S., Shibuya, N., Joosten, J., Thomma, B. (2010). Conserved Fungal LysM Effector Ecp6 Prevents Chitin-Triggered Immunity in Plants. Science, 329(5994), 953–955. doi:10.1126/science.1190859
- Deshmukh, S., Rai, M. (2005). Biodiversity of fungi: their role in human life. India. Science Pub Inc.https://cast.arizona.edu/mycoherb/arnoldlab/Arnold2005.chapter
- Djamei, A., Schipper, K., Rabe, F., Ghosh, A., Vincon, V., Kahnt, J., Osorio, S., Tohge, T., Fernie, A., Feussner, I., Feussner, K., Meinicke, P., Stierhof, Y., Schwarz, H., Macek, B., Mann, Ma., Kahmann, R. (2011). Metabolic priming by a secreted fungal effector. Nature, 478(7369), 395–398. doi:10.1038/nature10454
- Dodds, P & Thrall P. 2009. Recognition events and host-pathogen co-evolution in gene-for-gene resistance to flax rust. FunctionalPlant Biology 36: 395–408.
- Doehlemann, G., & Hemetsberger, C. (2013). Apoplastic immunity and its suppression by filamentous plant pathogens. New Phytologist, 198(4), 1001–1016. doi:10.1111/nph.12277
- Doehlemann, H and Hemetsberger, C. (2013) Apoplastic immunity and its suppression by filamentous plant pathogens. New Phytologist 198: 1001–1016
- Dohm, J., Peters, P., Stralis-Pavese, N., Himmelbauer, H. (2020). Benchmarking of long-read correction methods. NAR Genomics and Bioinformatics, 2(2), lqaa037–. doi:10.1093/nargab/lqaa037
- Dong, S., Raffaele, D., Kamoun S. (2015). The two-speed genomes of filamentous pathogens: waltz with plants Current Opinion in Genetics & Development, 35:57–65
- Duplessis, S., Cuomo, A., Lin, Y., Aerts, A., Tisserant, E., Veneault-Fourrey, C., Joly, D. L., Hacquard, S., Amselem, J., Cantarel, B. L., Chiu, R., et al. (2011). Obligate biotrophy features unraveled by the genomic analysis of rust fungi. Proceedings of the National Academy of Sciences, 108(22), 9166–9171. doi:10.1073/pnas.1019315108
- Edathil, T. (1986). South American leaf blight-A potential threat to the natural rubber industry in Asia and Africa. Tropical Pest Management, 32(4), 296–303. doi:10.1080/09670878609371083
- Eddy, R. (2009). A new generation of homology search tools based on probabilistic inference. Genome Inform., 23, 205–211.
- Eid, J. et al. Real-Time DNA Sequencing from Single Polymerase Molecules. Science 323, 133–138 (2009).
- Ejigu, F., & Jung, J. (2020). Review on the Computational Genome Annotation of Sequences Obtained by Next-Generation Sequencing. Biology. 2020; 9(9):295. https://doi.org/10.3390/biology9090295
- Ejigu, G & Jung, J. (2020).Review on the Computational Genome Annotation of Sequences Obtained by Next-Generation Sequencing. Biology 2020, 9, 295; doi:10.3390/biology9090295
- Emms, D., & Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol, 16, 157.

- Engelsdorf, T., Will, C., Hofmann, J., Schmitt,C., Merritt, B., Rieger, L., Frenger, M Marschall, A., Franke, R., Pattathil, R., Voll, L. (2017) Cell wall composition and penetration resistance against the fungal pathogen Colletotrichum higginsianum are affected by impaired starch turnover in Arabidopsis mutants. Journal of Experimental Botany, Vol. 68, No. 3 pp. 701–713, 2017
- Epstein, L., & Nicholson, R. (2006). Adhesion and Adhesives of Fungi and Oomycetes. Biological Adhesives (Vol. 107, pp. 25–55). Cham: Springer International Publishing.
- Ericsson, O., Hawksworth, D. (1993). Outline of the ascomycetes-1993. Syst Ascomycetum 12: 51–257
- Esse, H. P. van, Bolton, M. D., Stergiopoulos, I., Wit, P. J. G. M. de, & Thomma, B. P. H. J. (2007, agosto 8). The Chitin-Binding *Cladosporium fulvum* Effector Protein Avr4 Is a Virulence Factor (world) [Research-article]. Http://Dx.Doi.Org/10.1094/MPMI-20-9-1092; The American Phytopathological Society. https://doi.org/10.1094/MPMI-20-9-1092
- Feschotte, C, Swamy, L, Wessler. (2003). SR: Genome-wide analysis of mariner-like transposable elements in rice reveals complex relationships with Stowaway MITEs. Genetics 2003,163:747-758.
- Feschotte, C., & Mouchès, C. (2000) Evidence that a Family of Miniature Inverted-Repeat Transposable Elements (MITEs) from the Arabidopsis thaliana Genome Has Arisen from a pogo-like DNA Transposon, Molecular Biology and Evolution, Volume 17, Issue 5, May 2000, Pages 730–737, https://doi.org/10.1093/oxfordjournals.molbev.a026351
- Feschotte, N., Zhang, X., & Wessler, R. (2004). Using rice to understand the origin and amplification of miniature inverted repeat transposable elements (MITEs). Current Opinion in Plant Biology, 7(2), 115–119. doi:10.1016/j.pbi.2004.01.004
- Feschotte C, Mouche` s C: Evidence that a family of miniature inverted-repeat transposable elements (MITEs) from the Arabidopsis thaliana genome has arisen from a pogo-like DNA transposon. Mol Biol Evol 2000, 17:730-737.
- Fischer, M. (2003). The Lipase Engineering Database: a navigation and analysis tool for protein families. Nucleic Acids Research, 31(1), 319–321. doi:10.1093/nar/gkg015
- Flor, H. H. (1971). Current Status of the Gene-For-Gene Concept. Annual Review of Phytopathology, 9, 275- 296.
- Fouché, S., Plissonneau, C., & Croll, D. (2018). The birth and death of effectors in rapidly evolving filamentous pathogen genomes. Current Opinion in Microbiology, 46, 34–42. doi:10.1016/j.mib.2018.01.020
- Francois Bucchini, Andrea Del Cortona, Łukasz Kreft, Alexander Botzki, Michiel Van Bel, Klaas Vandepoele. (2020). TRAPID 2.0: a web application for taxonomic and functional analysis of de novo transcriptomes, doi: https://doi.org/10.1101/2020.10.19.345835
- Frantzeskakis, L., Kracher, B., Kusch, S. et al. (2018). Signatures of host specialization and a recent transposable element burst in the dynamic one-speed genome of the fungal barley powdery mildew pathogen. BMC Genomics 19, 381 (2018). https://doi.org/10.1186/s12864-018-4750-6
- Freeman, B.C. and G.A. Beattie. (2008). An Overview of Plant Defenses against Pathogens and Herbivores. The Plant Health Instructor. DOI: 10.1094/PHI-I-2008-0226-01
- Frey, K., & Pucker, B. (2020). Animal, Fungi, and Plant Genome Sequences Harbor Different Non-Canonical Splice Sites. Cells, 9(2), 458. doi:10.3390/cells902045

References

- Furtado E.L., de Jesus Junior W.C., Moraes W.B. (2020) Forest Diseases in Brazil: Status and Management. In: Estay S. (eds) Forest Pest and Disease Management in Latin America. Springer, Cham. https://doi.org/10.1007/978-3-030-35143-4_14
- Galagan, J. (2005). Genomics of the fungal kingdom: Insights into eukaryotic biology. Genome Research, 15(12), 1620–1631. doi:10.1101/gr.3767105
- Gallone, B., Steensels, J., Prahl, T., Soriaga, L., Saels, V., Herrera-Malaver, B.,Verstrepen, K. J. (2016). Domestication and divergence of Saccharomyces cerevisiae beer yeasts. Cell, 166, 1397–1410.e16. https://doi.org/10.1016/j.cell.2016.08.020
- Gao, D., Li, Y., Do, K., Abernathy, B., Jackson, A. (2016). Landscape and evolutionary dynamics of terminal repeat retrotransposons in miniature in plant genomes. Genome Biol. 2016;17(1):7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26781660
- Garcia, D., Sanier, C., Macheix, J., D'Auzac, J. (1995). Accumulation of scopoletin in *Hevea brasiliensis* infected by *Microcyclus ulei* (P. Henn.) V. ARX and evaluation of its fungitoxicity for three leaf pathogens of rubber tree. *, 47(4), 0–223.* doi:10.1006/pmpp.1995.1053
- García, I., Aristizábal, F., Montoya, D. A review of the Microcyclus ulei Ascomycetes fungus causative agent of South American rubber-leaf blight. (2006). Rev. Colomb. Biotecnol. Vol. VIII N° 2 Diciembre 2006 50-59
- García, R. I. A. (2012). Estudio de la interacción planta – patógeno en clones comerciales de Hevea brasiliensis presentes en jardines clonales de Colombia, susceptibles y resistentes al mal suramericano de la hoja del caucho [Thesis - Doctorado]. Universidad Nacional de Colombia.
- Garnica, D., Nemri, A., Upadhyaya, N., Rathjen, J., Dodds, P. (2014). The ins and outs of rust haustoria. PLoS Pathogens 10: e1004329
- Gasparatto, L., Figueiredo, A., Rezende, J.C., Ferreira, F.A. (1997). Doenças da Seringueira no Brasil. Empresa Brasilera de Pesquisa Agropecuarias. 39-41. EMBRAPA.
- Gasparotto, L., Lieberei, R., Trindade, D. (1984). In vitro conidia germination of Microcyclus ulei and its sensitivity to fungicides. Fitopatologia Brasileira. 9: 505-511.
- Geoghegan, I., Steinberg, G., Gurr, S. (2017) The Role of the Fungal Cell Wall in the Infection of Plants, Undefine.
- 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.
- Gomez, C., Murua, A., Garrido, M., Gonzalez, P., Mongue, R., Barber, D., Palacios, L., & Diáz, A. (2014). Alt a 1 from Alternaria interacts with PR5 thaumatin-like proteins. FEBS Letters, 588(9), 1501-1508. https://doi.org/10.1016/j.febslet.2014.02.044
- Gonçalves, P., Ortolani, A., Cardoso, M. (1997). Melhoramento genético da seringueira: uma revisão. Instituto Agronômico, Campinas (Brazil): Instituto Agronômico, Campinas (Brazil);.
- Gonzalez, S., Oggenffus, U., García, I., Aristizábal, F., Croll, D., & Riaño-Pachon, D. M. (2020, julio 13). The Pseudocercospora genome assembly reveals a significant size expansion mediated by specific transposable elements. ISBM-2020.
- Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next- generation sequencing technologies. Nature Publishing Group, 17(6), 333–351.
- Goodwin, S. B., Ben M'Barek, S., Dhillon, B., Wittenberg, A. H. J., Crane, C. F., Hane, J. K., Foster, A. J., Van der Lee, T. A. J., Grimwood, J., Aerts. Et al. (2011). Finished genome of the fungal wheat pathogen Mycosphaerella graminicola reveals

dispensome structure, chromosome plasticity, and stealth pathogenesis. PLoS Genet. 7:e1002070.

- Govers, F., & Gijzen, M. (2006). Phytophthora Genomics: The Plant Destroyers' Genome Decoded. Molecular Plant-Microbe Interactions, 19(12), 1295–1301.
- Graovac, M. & Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc. Bioinformatics 25, 4.10.1–4.10.14.
- Grigoriev, IV., Nikitin, R., Haridas, S., Kuo, A., Ohm, R., Otillar, R., et al. MycoCosm portal: gearing up for 1000 fungal genomes. Nucleic Acids Res. 2014;42 (Database issue): D699–704.
- Gui, J., Zhang, Q., Zhang,D., Zhou, L., Short, G., Wang, J., Ma, F., Li, G., Kong, Q., Wang, L., Wang, D., Li, Y., Subbarao, V., Chen, Y., & Dai, F. (2017). A Verticillium dahliae Extracellular Cutinase Modulates Plant Immune Responses. Molecular Plant-Microbe Interactions. https://doi.org/10.1094/MPMI-06-17-0136-R
- Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics, 29(8), 1072–1075. doi:10.1093/bioinformatics/btt086
- Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics (Oxford, England), 29(8), 1072–1075. https://doi.org/10.1093/bioinformatics/btt086
- Guyot, J., & Le Guen, V. (2017). A review of a century of studies on South American Leaf Blight of the rubber tree. Plant Disease, PDIS-04-17-0592-FE–. doi:10.1094/PDIS-04-17-0592-FE
- Haas, B., Papanicolaou, A., Yassour, M. et al. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 8, 1494–1512 (2013). https://doi.org/10.1038/nprot.2013.084
- Haas, B. J., Zeng, Q., Pearson, M. D., Cuomo, C. A., & Wortman, J. R. (2011). Approaches to Fungal Genome Annotation. Mycology, 2(3), 118–141. https://doi.org/10.1080/21501203.2011.606851
- Haas, H., Eisendle, M., Turgeon, G. (2008). Siderophores in Fungal Physiology and Virulence. , 46(1), 149–187. doi:10.1146/annurev.phyto.45.062806.094338
- Haas et al. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biology 2008, 9:R7doi:10.1186/gb-2008-9-1-r7.
- Haft, H., Selengut, D., Richter, A., Harkins, D., Basu, K., Beck, E. (2013). TIGRFAMs and Genome Properties in 2013. Nucleic Acids Res. 2013 Jan;41(Database issue):D387-95. doi: 10.1093/nar/gks1234. Epub 2012 Nov 28. PMID: 23197656; PMCID: PMC3531188.
- Hall, B., DeRego, T., & Geib, S. (2014). GAG: the Genome Annotation Generator (Version 1.0). Available from http://genomeannotation.github.io/GAG.
- Han, J., Pluhackova, K., & Böckmann, R. A. (2017). The Multifaceted Role of SNARE Proteins in Membrane Fusion. Frontiers in Physiology, 8. https://doi.org/10.3389/fphys.2017.00005
- Han, J., Pluhackova, K., & Böckmann, R. A. (2017). The Multifaceted Role of SNARE Proteins in Membrane Fusion. Frontiers in Physiology, 8. https://doi.org/10.3389/fphys.2017.00005
- Hane, J., Paxman, J., Jones, D., Oliver, R & de Wit, Pierre. (2020) “CATAStrophy,” a Genome-Informed Trophic Classification of Filamentous Plant Pathogens – How Many Different Types of Filamentous Plant Pathogens Are There?. Frontiers in Microbiology (10): 3088.p.1-12

References

- Haridas S, Albert R, Binder M, et al. (2020). 101 Dothideomycetes Genomes: a test case for predicting lifestyles and emergence of pathogens. Studies in Mycology. S0166061620300038–. doi:10.1016/j.simyco.2020.01.003
- Hashim, I., Chee, K., Dunkan, E. (1978) Reaction of *Hevea* Leaves to Infection with *Microcyclus ulei.* Rubb. Res. Imt. Malaysia*, 26(2), 67-7*
- Hashim, I., Chee, K. H., & Duncan, E. J. (1978). Reaction of Hevea leaves to infection with Microcyclus ulei. 26(2), 67-75.
- Hayashi, Y. Production of natural rubber from Para rubber tree. (2009) Plant Biotechnology 26, 67–70
- Heng, Li. (2018) Minimap2: pairwise alignment for nucleotide sequences, Bioinformatics, Volume 34, Issue 18, 15 September 2018, Pages 3094–3100, https://doi.org/10.1093/bioinformatics/bty191
- Hennings P (1904) Uber die auf Hevea –arten bisher beobachteten parasitischen pilze. Notizbl bot Gart Mus Berl 4: 133–139.
- Hernández, M., Pérez, L., Niño, G., & Mora, M. (2017). Fungal Strategies to Evade the Host Immune Recognition. Journal of fungi (Basel, Switzerland), 3(4), 51. https://doi.org/10.3390/jof3040051
- Hoff, K., & Stanke, M. (2015). Current methods for automated annotation of protein-coding genes. Current Opinion in Insect Science, 7, 8–14. doi:10.1016/j.cois.2015.02.008
- Hoff, K.J., Lange, S., Lomsadze, A., Borodovsky, M., Stanke, M. 2015. BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics 32: 767–769.
- Holloway, P. (1982). Structure and histochemistry of plant cuticular membranes: An overview. Pages 1-32 in: The Plant Cuticle. D. F. Cutler, K. L. Alvin. and C. E. Price, eds. Academic Press, New York.
- Hora, D., de Macedo, M., Barreto, W., Evans, C., Mattos, C., Maffia, A., & Mizubuti, E. S. G. (2014). Erasing the Past: A New Identity for the Damoclean Pathogen Causing South American Leaf Blight of Rubber. PLoS ONE, 9(8), e104750–12.
- Horbachab, R., Navarro, A., Wolfgang, B., Deising, B. (2011). When and how to kill a plant cell: Infection strategies of plant pathogenic fungi. Journal of Plant Physiology 168 (2011) 51–62
- Horton, P., Park, K., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C., Nakai, K. (2007) WoLF PSORT: protein localization predictor. Nucleic Acids Res. Jul; 35(Web Server issue): Pages 585–587.
- Houston, K., Tucker, M., Chowdhury, J., Shirley, N., Little, A. (2016). The Plant Cell Wall: A Complex and Dynamic Structure As Revealed by the Responses of Genes under Stress Conditions. Frontiers in Plant Science, 7(), –. doi:10.3389/fpls.2016.00984
- Howard, R., Ferrari, M., Roach, D., Money, N. (1991). Penetration of hard substrates by a fungus employing enormous turgor pressures. Proceedings of the National Academy of Sciences, USA 88: 11281– 11284.
- Hunter S. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014 May 1;30(9):1236-40. doi: 10.1093/bioinformatics/btu031. Epub 2014 Jan 21. PMID: 24451626; PMCID: PMC3998142.
- Hurtado, U., García, IA., Restrepo, S., Aristizábal, F., Montoya D. (2015) Assembly and Analysis of Differential Transcriptome Responses of Hevea brasiliensis on Interaction with Microcyclus ulei. PLoS ONE 10(8): e0134837. doi:10.1371/journal. pone.0134837
- International Rubber Study Group (IRSG). Rubber Statistical Bulletin. Sri Lanka; 2019.

- Isaza, R., Diaz, C., Dhillon, B., Aerts, A., Carlier, J., Crane, CF. (2016) Combating a Global Threat to a Clonal Crop: Banana Black Sigatoka Pathogen Pseudocercospora fijiensis (Synonym Mycosphaerella fijiensis) Genomes Reveal Clues for Disease Control. PLoS Genetic 12(8) 1-36
- Jacobson, E. (2000) pathogenic roles for fungal melanins. clinical microbiology p. 708–717 vol. 13, no. 4 0893-8512
- Jaswal, R., Kiran, K., Rajarammohan, S., Dubey, H., Singh, P. K., Sharma, Y., Deshmukh, R., Sonah, H., Gupta, N., & Sharma, T. R. (2020). Effector Biology of Biotrophic Plant Fungal Pathogens: Current Advances and Future Prospects. Microbiological Research, 241, 126567. https://doi.org/10.1016/j.micres.2020.126567
- Jiang N., Feschotte, C., Zhang, X., Wessler, S. (2004). Using rice to understand the origin and amplification of miniature inverted repeat transposable elements (MITEs). , 7(2), 115–119. doi:10.1016/j.pbi.2004.01.004
- Johnson, L. (2008). Iron and siderophores in fungal–host interactions. , 112(2), 170–183. doi:10.1016/j.mycres.2007.11.012
- Jones, J & Dangl, J. (2006). The plant immune system. Nature, 444, 323-329.
- Jones, P., Binns, D., Chang, H., Fraser, M., Li, W., McAnulla, C., MacWillian, H., Maslen, J., Mitchel, A., Nuka, G. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics. 30(9): Pages 1236–40.
- Jones, P., Binns, D., Chang, Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., et al. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014 May 1;30(9):1236-40. doi: 10.1093/bioinformatics/btu031. Epub 2014 Jan 21. PMID: 24451626; PMCID: PMC3998142.
- Jun-Ma, L., Fedorova, N. (2010). A practical guide to fungal genome projects: strategy, technology, cost and completion. Mycology, 1(1), 9–24. doi:10.1080/21501201003680943
- Junqueira, N., Chaves, G., Zambolim, L., Alfenas, A., Gasparotto, L. (1988). Reacao de clones de seringueira a vários isolados de Microcyclus ulei. Pesquisa Agropecuária Brasileira 23: 877–893
- Junqueira, N., Chaves, G., Zambolim, L., Alfenas, A., Gasparotto, L. (1988). Reacao de clones de seringueira a vários isolados de Microcyclus ulei. Pesquisa Agropecuária Brasileira 23: 877–893
- Junqueira, N., Kalili, F,. Araujo, A., 1992. Genètica da resistência da seringueira ao Microcyclus ulei. Fitopatologia Brasileira. 14 (2): 149.
- Kalpani, K., Withange, S., Palihakkara, R. (2020) Selection of Superior Genotypes at Early Stage of the Rubber (Hevea brasiliensis) Breeding Cycle. ISSN: 2349-8889 Volume-7, Issue-4
- Kamoun P. (2014) How do filamentous pathogens deliver effector proteins into plant cells? PLoS Biol. 2014, 12
- Katoh, K., Kuma, K., Toh, H., Miyata, T. (2005). MAFFT version 5: Improvement in accuracy of multiple sequences alignment. Nucleic Acids Res 33: Pages 511-518.
- Keilwagen, J., Hartung, F., Paulini, M. et al. Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. BMC Bioinformatics 19, 189 (2018). https://doi.org/10.1186/s12859-018-2203-5
- Keller, N., Turner, G., Bennett, J. (2005). Fungal secondary metabolism — from biochemistry to genomics. , 3(12), 937–947. doi:10.1038/nrmicro1286
- Keller, O., Odronitz, F., Stanke, M., Kollmar, M., & Waack, S. (2008). Scipio: Using protein sequences to determine the precise exon/intron structures of genes and

their orthologs in closely related species. BMC Bioinformatics, 9(1), 278. doi:10.1186/1471-2105-9-278

- Kempken, F., & Kück, U. (1998). Transposons in filamentous fungi--facts and perspectives. Bioessays. 1998 Aug;20(8):652-9. Doi: 10.1002/(SICI)1521-1878(199808)20:8<652::AID-BIES8>3.0.CO;2-K. PMID: 9841641.
- Khaldi, N., Seifuddin, F., Turner, G., Haft, D., Nierman, W., Wolfe, K., Fedorova, N (2010). SMURF: Genomic mapping of fungal secondary metabolite clusters. , 47(9), 0–741. doi:10.1016/j.fgb.2010.06.003
- Kim, D., Langmead, B., Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. Nat Methods 12:357–360. https://doi.org/10 .1038/nmeth.3317.
- Kingston, E. (2000). Preparation and Analysis of RNA. UNIT 4.3: Phenol/SDS Method for Plant RNA Preparation. In J. Whiley & &. Sons (Eds) Current Protocols in Molecular Biology (pp. 4.3.1-4.3.4).
- Kleemann, J., Rincon, J., Takahara, H., Neumann, U., van Themaat, E., van der Does, C., O'Connell, R. (2012). Sequential Delivery of Host-Induced Virulence Effectors by Appressoria and Intracellular Hyphae of the Phytopathogen Colletotrichum higginsianum. PLoS Pathogens, 8(4), e1002643. doi:10.1371/journal.ppat.1002643
- Kleemann, Jochen; Rincon-Rivera, Linda J.; Takahara, Hiroyuki; Neumann, Ulla; van Themaat, Emiel Ver Loren; van der Does, H. Charlotte; Hacquard, Stéphane; Stüber, Kurt; Will, Isa; Schmalenbach, Wolfgang; Schmelzer, Elmon; O'Connell, Richard J.; Howlett, Barbara J. (2012). Sequential Delivery of Host-Induced Virulence Effectors by Appressoria and Intracellular Hyphae of the Phytopathogen Colletotrichum higginsianum. PLoS Pathogens, 8(4), e1002643–. doi:10.1371/journal.ppat.1002643
- Knogge, W. (1998). Fungal pathogenicity. Current Opinion in Plant Biology, 1(4), 324–328.
- Kolattukudy, P. (1985). Enzymatic penetration of the plant cuticle by fungal pathogens. Annu. Rev. Phytopathol. 23:223-250.
- Kombrink,A., Rovenich, A., Shi-Kunne, X., Rojas,Eduardo., Van Den Berg G et al. (2017) Verticillium dahliae LysM effectors differentially contribute to virulence on plant hosts Molecular plant pathology 18(4), 596–608
- Koonin, V., Galperin, Y. (2003). Sequence - Evolution - Function: Computational Approaches in Comparative Genomics. Boston: Kluwer Academic; Chapter 5, Genome Annotation and Analysis. Available from: https://www.ncbi.nlm.nih.gov/books/NBK20253/
- Koop, D. M., Rio, M., Sabau, X., Cardoso, S. E. A., Cazevieille, C., Leclercq, J., & Garcia, D. (2016). Expression analysis of ROS producing and scavenging enzyme-encoding genes in rubber tree infected by Pseudocercospora ulei. Plant Physiology and Biochemistry, 104, 188-199.
- Koren, S., & Phillippy, A. (2015). One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. Current Opinion in Microbiology 2015, 23:110–120
- Koren, S., Brian, P., Walenz Berlin, K.,Miller, J., Bergman, N., Phillip, A. (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation
- Krishnan, P., Ma, X., McDonald, B. A., & Brunner, P. C. (2018). Widespread signatures of selection for secreted peptidases in a fungal plant pathogen. BMC Evolutionary Biology, 18(1). doi:10.1186/s12862-018-1123-3
- Kroken, S., Glass, L., Taylor, W., Yoder, C., Turgeon, G. (2003). Phylogenomic analysis of type I polyketide synthase genes in pathogenic and saprobic

ascomycetes. Proceedings of the NationalAcademyof Sciences. 2003; 100(26):15670–5. https://doi.org/10.1073/pnas.2532165100 PMID: 14676319

- Kuana, T., Zhaia, Y., Maa, W. (2016) Small RNAs regulate plant responses to filamentous pathogens. Cell & Developmental Biology 56 (2016) 190–200
- Kubicek, P., Starr, L., & Glass, N. L. (2014). Plant cell wall-degrading enzymes and their secretion in plant-pathogenic fungi. Annual Review of Phytopathology, 52, 427-451. https://doi.org/10.1146/annurev-phyto-102313-045831
- Kück, P & Longo, G. (2014). FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. Front Zool 11, Page 81.
- Kupfer, M., Drabenstot, D., Buchanan, L., Lai, H., Zhu, H., Dyer, W., Murphy, W. (2004). Introns and Splicing Elements of Five Diverse Fungi. Eukaryotic Cell, 3(5), 1088–1100. doi:10.1128/ec.3.5.1088-1100.2004
- Laehnemann,D., Borkhardt, A., & McHardy, C. (2016) . Denoising DNA deep sequencing data-high-throughput sequencing errors and their correction. Brief. Bioinformatics, 17, 154–179.
- Laetsch, R., & Blaxter L. (2017). BlobTools: Interrogation of genome assemblies [version 1; peer review: 2 approved with reservations]. F1000Research, 6:1287
- Langfelder, K., Streibel, M., Jahn, B., Haase, G., Brakhage, A. (2003) Biosynthesis of fungal melanins and their importance for human pathogenic fungi. Fungal Genetics and Biology 38: 143–158.
- Le Guen, V.,, Rodier-Goud, M., Troispoux,V., Xiong, ., Brottier, P., Billot, P., Seguin, (2004) Characterization of polymorphic microsatellite markers for Microcyclus ulei, causal agent of South American leaf blight of rubber trees.Molecular Ecology Notes (2004) 4, 122–124
- Le Guen, V., Garcia, D., Mattos. C., Doaré, F., Lespinasse, D. (2006) Bypassing of a polygenic Microcyclus ulei resistance in rubber tree, analyzed by QTL detection. New Phytologist;173(2): 335 –345.
- Le Guen, V., Garcia, D., Mattos. C., Doaré, F., Lespinasse, D. (2008). Bypassing of a polygenic Microcyclus ulei resistance in rubber tree, analyzed by QTL detection. New Phytologist;173(2): 335 –345
- Le Guen, V., Guyot, J., Mattos, C. R. R., Seguin, M. & Garcia, D. (2008). Long-lasting rubber tree resistance to Microcyclus ulei characterized by reduced conidial emission and absence of teleomorph. Crop Protection, 27, 1498-1503
- Le Guen, V., Lespinasse, D., Oliver, G., Rodier-Goud, M., Pinard, F. & Seguin, M. (2003). Molecular mapping of genes conferring field resistance to South American Leaf Blight (Microcyclus ulei) in rubber tree. TAG Theoretical and Applied Genetics, 108, 160-167.
- Lelwala, R., Korhonen, P., Young, N., Scott, J., Ades, P., Gasser, R., Taylor, P.,Coleman, C. (2019). Comparative genome analysis indicates high evolutionary potential of pathogenicity genes in Colletotrichum tanaceti. PLOS ONE, 14(5), e0212248–. doi:10.1371/journal.pone.0212248
- Lelwala, R., Korhonen, P., Young, N., Scott, J., Ades, P., Gasser, R., Taylor, P., Coleman, C. (2019). Comparative genome analysis indicates high evolutionary potential of pathogenicity genes in Colletotrichum tanaceti. PLOS ONE, 14(5), e0212248–. doi:10.1371/journal.pone.0212248
- Lespinasse, D., Grivet, L., Troispoux, V., Rodier-Goud, M., Pinard, F. & Seguin, M. (2000a). Identification of QTLs involved in the resistance to South American leaf blight (Microcyclus ulei) in the rubber tree. TAG Theoretical and Applied Genetics, 100, 975-984.

References

- Lespinasse, D., Rodier-Goud, M., Grivet, L., Leconte, A., Legnate, H. & Seguin, M. (2000b). A saturated genetic linkage map of rubber tree (Hevea spp.) based on RFLP, AFLP, microsatellite, and isozyme markers. Theoretical and Applied Genetics, 100, 127-138.
- Letunic, I., Doerks, T., Bork, P. (2012). SMART 7: recent updates to the protein domain annotation resource. Nucleic Acids Res. 2012 Jan;40(Database issue):D302-5. doi: 10.1093/nar/gkr931. Epub 2011 Nov 3. PMID: 22053084; PMCID: PMC3245027.
- Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., … Wang, J. (2009). De novo assembly of human genomes with massively parallel short read sequencing. Genome *Research, 20(2), 265–272.* doi:10.1101/gr.097261.109
- Li, X., Griffin, K., Langeveld, S., Frommhagen, M., Underlin, E., Kabel, M., de Vries, R., Dilokpimol, A. (2020). Functional Validation of Two Fungal Subfamilies in Carbohydrate Esterase Family 1 by Biochemical Characterization of Esterases From Uncharacterized Branches. Frontiers in Bioengineering and Biotechnology, 8, 694–. doi:10.3389/fbioe.2020.00694
- Li, X., Griffin, K., Langeveld, S., Frommhagen, M., Underlin, E. N., Kabel, M. A., … Dilokpimol, A. (2020). Functional Validation of Two Fungal Subfamilies in Carbohydrate Esterase Family 1 by Biochemical Characterization of Esterases From Uncharacterized Branches. Frontiers in Bioengineering and Biotechnology, 8. doi:10.3389/fbioe.2020.00694
- Li. Z, Chen. Y, Mu. D, Yuan. J, Shi. Y, Zhang. H, Gan. J, Li. N, Hu. X, Liu. B, Yang, B and Fan. W. (2011) Comparison of the two major classes of assembly algorithms: overlap layout consensus and de-bruijn-graph. Briefings in functional genomics. Vol 11. NO 1. 25- 37
- Lieberei, R., Schrader, A., Biehl, B., Chee, K. (1983). Effect of cyanide on Microcyclus ulei cultures. Journal of the Rubber Research, Institute of Malaysia, 31, 227-235. Babraham Bioinformatics, Cambridge, UK.
- Lieberei, R. (2006). Physiological characteristics of Microcyclus ulei (P. Henn.) V.ARX. - A fungal pathogen of the cyanogenic host Hevea brasiliensis. Journal of Applied Botany and Food Quality, 80, 63-68.
- Lieberei, R. (2007). South American leaf blight of the rubber tree (Hevea spp.): new steps in plant domestication using physiological features and molecular markers. Annals of Botany, 100(6), 1125–1142
- Lin, Y., Yuan, J., Kolmogorov, M., Shen, M. W., Chaisson, M., & Pevzner, P. A. (2016). Assembly of long error-prone reads using de Bruijn graphs. Proceedings of the National Academy of Sciences*, 113(52), E8396–E8405.* doi:10.1073/pnas.1604560113
- Linder M., Szilvay, G., Nakari-Setälä, T., Penttilä, M. (2005). Hydrophobins: the protein-amphiphiles of filamentous fungi. 29(5), 877–896. doi:10.1016/j.femsre.2005.01.004
- Liu, B., Shi, Y., Yuan, J., Hu, X., Zhang, H., Li, N., Li, Z., Chen, Y., Mu. et al. (2012) Estimation of genomic characteristics by analyzing *kmer* frequency in de novo genome projects.
- Liu, Z., Zhang, Z., Faris, J., Oliver R.P, Syme R, et al. (2012) The Cysteine Rich Necrotrophic Effector SnTox1 Produced by Stagonospora nodorum Triggers Susceptibility of Wheat Lines Harboring Snn1. PLoS Pathog 8(1)
- Loehrer, M. Vogel, A. Huettel, B. Reinhardt, R. Benes, V. Duplessis, S. Usadel, B. Schaffrath, U.(2014). On the current status of Phakopsora pachyrhizi genome sequencing.Frontiers in Plant Science (5): 1-5
- Lombard, V., Golaconda, H., Drula, El., Coutinho, P., Henrissat, B. (2014). The carbohydrate-active enzymes database (CAZy) in 2013, Nucleic Acids Research,

Volume 42, Issue D1, 1 January, Pages D490–D495, https://doi.org/10.1093/nar/gkt1178

- Lombard,V., Golaconda Ramulu,H., Drula,E., Coutinho,P.M. and Henrissat,B. (2014) The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res., 42, D490–D495.
- Lomsadze, A., Burns, P. D., & Borodovsky, M. (2014). Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithms. Nucleic acids research, 42(15), e119. https://doi.org/10.1093/nar/gku557
- Lomsadze A. et al. . (2014) Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithms. Nucleic Acids Res., 42, e119.
- Lopez, D., Ribeiro, S., Label, P., Fumanal, B., Venisse, J., Kohler, A., de Oliveira, R., Labutti, K., Lipzen, A., Lail, Ka., Bauer, D., Ohm, R., Barry, K., Spatafora, J., Grigoriev, I., Martin, F., Pujade-Renaud, V. (2018). Genome-Wide Analysis of Corynespora cassiicola Leaf Fall Disease Putative Effectors. Frontiers in Microbiology*,* 9(), 276–. doi:10.3389/fmicb.2018.00276
- Lo Presti, L., & Kahmann, R. (2017). How filamentous plant pathogen effectors are translocated to host cells. Current Opinion in Plant Biology, 38, 19–24. doi:10.1016/j.pbi.2017.04.005
- Lo Presti, L., Lanver, D., Schweizer, G., Tanaka, S., Liang, L., Tollot, M., Zuccaro, A., Reissmann, S., Kahmann, R. (2015). Fungal Effectors and Plant Susceptibility. Annual Review of Plant Biology, 66(1), 513–545. doi:10.1146/annurev-arplant-043014-114623
- Lumbsch TH, Huhndorf SM. (2007). "Outline of Ascomycota – 2007". Myconet. Chicago, USA: The Field Museum, Department of Botany. 13: 1–58.
- Luo et al., (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler GigaScience, Vol. 1, No. 1, 18, doi:10.1186/2047-217x-1-18
- Lyu, X., Shen, C., Fu, Y., Xie, J., Jiang, D., Li, G., & Cheng, J. (2015). Comparative genomic and transcriptional analyses of the carbohydrate-active enzymes and secretomes of phytopathogenic fungi reveal their significant roles during infection and development. Scientific Reports, 5(1). doi:10.1038/srep15565
- Lyu, X., Shen, C., Fu, Y. et al. Comparative genomic and transcriptional analyses of the carbohydrate-active enzymes and secretomes of phytopathogenic fungi reveal their significant roles during infection and development. Sci Rep 5, 15565 (2015). https://doi.org/10.1038/srep15565
- Ma, B., Li, T., Xiang, Z., et al. (2015). TE db a collective resource for mulberry transposable elements. Database. 2015;2015 Available from: https://academic.oup.com/database/article/doi/10.1093/database/bav004/2433136
- Macheleidt, J., Mattern, D., Fischer, J., Netzker, T., Weber, J., Schroeckh, V., Valiante, V.,Brakhage, A. (2016). Regulation and Role of Fungal Secondary Metabolites.Annu. Rev. Genet. 2016. 50:16.1–16.22
- Magi, A., Giusti, B. & Tattini, L. Characterization of MinION nanopore data for resequencing analyses. Brief. Bioinforma. 18, 940–953 (2016).
- Mao, H., Wang, H. (2017). SINE_scan: an efficient tool to discover short interspersed nuclear elements (SINEs) in large-scale genomic datasets. Bioinformatics. 2017;33(5):btw718. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28062442
- Marcais, G., Kingsford, C., (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27(6): 764-770

References

- Mattos, C., García, D., Pinard, F., Le Guen, V. (2003). Variabilidade de Isolados de Microcyclus ulei no Sudeste de Bahia. Fitopatologia Brasileira 28 (5): 502-507
- Mattos, C., García, D., Pinard, F., Le Guen, V. (2003). Variabilidade de Isolados de Microcyclus ulei no Sudeste de Bahia. Fitopatologia Brasileira 28 (5): 502-507
- Mendez, T. M. E. (2017). Identificación de moléculas candidatas a proteínas efectoras de Microcyclus ulei presentes en la interacción con *Hevea brasiliensis* [Maestria]. Universidad Nacional de Colombia.
- Mendgen, K., Hahn, M., & Deising, H. (1996). Morphogenesis and mechanisms of penetration by plant pathogenic fungi. Annual Review of Phytopathology, 34(1), 367–386.
- Menzel P., Ng L., Krogh A. (2016) Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat. Commun. 7:11257
- Mesarich, C. H., Ökmen, B., Rovenich, H., Griffiths, S. A., Wang, C., Jashni, M. K., Mihajlovski, A., Collemare, J., Hunziker, L., Deng, C. H., Burgt, A. van der, Beenen, H. G., Templeton, M. D., Bradshaw, R. E., & Wit, P. J. G. M. de. (2017). Specific hypersensitive response-associated recognition of new apoplastic effectors from Cladosporium fulvum in wild tomato. BioRxiv, 127746. https://doi.org/10.1101/127746
- Mikheyev, A., & Tin, Y. (2014). A first look at the Oxford Nanopore MinION sequencer. Molecular Ecology Resources, 14(6), 1097–1102. doi:10.1111/1755-0998.12324
- Miller, J., Koren, S., Sutton, G. (2010) Assembly algorithms for next-generation sequencing data. Genomics 2010;95: 315–27.
- Miller, R., Zhou, P, Mudge, J, Gurtowski J, Lee, H , Ramaraj, T, Walenz, P, Liu, J, Stupar, M, et al. (2017) Hybrid assembly with long and short reads improves discovery of gene family expansions. BMC Genomics 18:541
- Mohanta & Bae 2015.The diversity of fungal genome. Biological Procedures Online 17:8 DOI 10.1186/s12575-015-0020-z
- Mohanta & Bae 2015. The diversity of fungal genome. Biological Procedures Online 17:8 DOI 10.1186/s12575-015-0020-z
- Möller, M., Stukenbrock, E. (2017). Evolution and genome architecture in fungal plant pathogens. Nature Reviews (15):757-769
- Möller, M., Stukenbrock. E. (2017) Evolution and genome architecture in fungal plant pathogens. Nature reviews (15):757-769
- Mooibroek H, Cornish K (2000). Alternative sources of natural rubber. Appl. Microbiol. Biotechnol. 53:355-365.
- Mueller, O., Kahmann, R., Aguilar, G., Trejo, B., Wu, A., P. de Vries, P. (2008).The secretome of the maize pathogen Ustilago maydis. Fungal Genetics and Biology 45 S63–S70
- Mukund, K., Gupta, S., Prabhakar, K., Ranjekar. (2001). Differential Distribution of Simple Sequence Repeats in Eukaryotic Genome Sequences. Plant Molecular Biology 1–7.
- Muñoz, C., Vitte, C., Ross, J., Gout, S., Tenaillon, I. (2012). Using Nextgen Sequencing to Investigate Genome Size Variation and Transposable Element Content, Topics in Current Genetics 24,3-642-31842-93
- Muszewska, A., Hoffman, M., Grynberg, M. (2011) LTR Retrotransposons in Fungi. PLoS ONE 6(12): e29425. https://doi.org/10.1371/journal.pone.0029425
- Nakashima, C., Motohashi, K., Chen, C., Groenewald, J., Crous, P. (2016) Species diversity of Pseudocercospora from Far East Asia. Mycol Prog, 15, 1093–1117
- Niedringhaus, P., Milanova, D., Kerby, B., Snyder, A., Barron, E., (2011) Landscape of next-generation sequencing technologies Anal. Chem., 83 (2011), pp. 4327-4341

- Nielsen, H. (2017). Predicting Secretory Proteins with SignalP. In: Kihara D, editor. Protein Function Prediction: Methods and Protocols. New York, NY: Springer New York; Pages 59–73.
- Nietsch, R., Haas, J., Lai, A., Oehler, D., Mester, S., Frese, K., Sedaghat-Hamedani, F., Kayvanpour, E., Keller, A., Meder, B. (2016). The Role of Quality Control in Targeted Next-generation Sequencing Library Preparation.Genomics Proteomics Bioinformatics 14 (2016) 200–206
- Nilton, T., Junqueira, N., Lima, M., Gasparotto, L., Luis, A. (1992). Integrated control of rubber tree leaf Blight association between genetic resistente and chemical control. Pesq. Agropec. Bras. 27 (7): 1027-1034
- Oghenekome, U. (2004). Natural rubber, Hevea brasiliensis (Willd. ex A. Juss.) Mull. Arg, germplasm collection in the Amazon Basin, Brazil: A retrospective. Economic Botany, 58, 544-555
- O'Connell, R., Panstruga , R. (2006) Tête à tête inside a plant cell: establishing compatibility between plants and biotrophic fungi and oomycetes New Phytologist 171: 699–718
- Parisot, N., Vargas, C., Goubert, C., Baa-Puyoulet, P., Balmand,S., Beranger, L., Blanc, C., Bonnamour, A., Boulesteix, M,m Burlet, N., Calevro,F. (2021).The genome sequence of the cereal pest Sitophilus oryzae: an unprecedented transposable element content bioRxiv 2021.03.03.408021; doi: https://doi.org/10.1101/2021.03.03.408021
- Pascoal, A., Estevinho, L. M., Martins, I. M., & Choupina, A. B. (2018). Novel sources and functions of microbial lipases and their role in infection mechanisms. Physiological and Molecular Plant Pathology. doi:10.1016/j.pmpp.2018.08.003
- Pathak, D., Ngai, K., & Ollis, D. (1988). X-ray crystallographic structure of dienelactone hydrolase at 2.8 Å. Journal of Molecular Biology, 204(2), 435-445. https://doi.org/10.1016/0022-2836(88)90587-6
- Pertea, M., Kim, D., Pertea, M., Leek, T., Salzberg, L. (2016) Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown, Nature Protocols 11, 1650-1667, doi:10.1038/nprot.2016.095
- Pertea, M., Pertea, M., Antonescu, M., Chang, C., Mendell, T., Salzberg L. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads Nature Biotechnology, doi:10.1038/nbt.3122
- Pertea G and Pertea M. GFF Utilities: GffRead and GffCompare [version 1; peer review: 3 approved]. F1000Research 2020, 9:304 (https://doi.org/10.12688/f1000research.23297.1)
- Petrini, O. (1991) in Microbial Ecology of Leaves, eds. Andrews, J. H. & Hirano, S. S. Springer, New York, pp. 179–197.
- Pierleoni, A., Martelli, L., Casadio, R. (2008). PredGPI: a GPI-anchor predictor. BMC Bioinformatics. 2008;9(1):392. doi: 10.1186/1471-2105-9-392.
- Plissonneau, C., Benevenuto, J., Mohd-Assaad, N., Fouché, S., Hartmann, F. E., & Croll, D. (2017). Using Population and Comparative Genomics to Understand the Genetic Basis of Effector-Driven Fungal Pathogen Evolution. Frontiers in Plant Science, 8. doi:10.3389/fpls.2017.00119
- Priest, S., Yadav, V., & Heitman J. (2020). Advances in understanding the evolution of fungal genome architecture [version 1; peer review: 2 approved]. F1000Research, 9(Faculty Rev):776 (https://doi.org/10.12688/f1000research.25424.1)

References

- Priyadarshan, M., Gonçalves, P & Omokhafe, K. O. (2009). Breeding Hevea Rubber. Breeding Plantation Tree Crops: Tropical Species, 469–522. doi:10.1007/978-0-387-71201-7_13
- Pusztahelyi, T., Holb, J., Pócsi, I. (2015). Secondary metabolites in fungus-plant interactions. Frontiers in Plant Science, 6(), –. doi:10.3389/fpls.2015.00573
- Pusztahelyi, T. (2018). Chitin and chitin-related compounds in plant–fungal interactions. Mycology, (), 1–13. doi:10.1080/21501203.2018.1473299
- Quaedvlieg, W., Groenewald, J., Yáñez-Morales, M de Jesús., Crous, P. (2012). DNA barcoding of Mycosphaerella species of quarantine importance to Europe. Persoonia 29: 101–115
- Quinlan, A., & Hall, I. (2010) BEDTools: a flexible suite of utilities for comparing genomic features, Bioinformatics, Volume 26, Issue 6, 15 March 2010, Pages 841–842, https://doi.org/10.1093/bioinformatics/btq033
- Rahman, A., Usharraj, O., Misra, B. B., Thottathil, G. P., Jayasekaran, K., Feng, Y., et al. (2013). Draft genome sequence of the rubber tree Hevea brasiliensis. BMC Genomics, 14, 75–75.
- Rajarammohan, S., Paritosh, K., Pental, D., Kaur, J. (2019). Comparative genomics of Alternaria species provides insights into the pathogenic lifestyle of *Alternaria brassicae* a pathogen of the Brassicaceae family. BMC Genomics, 20(1), 1036–. doi:10.1186/s12864-019-6414-6
- Rang, F. J., Kloosterman, W. P. & Ridder, J. D. From squiggle to base-pair: computational approaches for improving nanopore sequencing read accuracy. Genome Biol. 19, 90 (2018).
- Rao, S., Sharda, S., Oddi, V., Nandineni, R. (2018). The Landscape of Repetitive Elements in the Refined Genome of Chilli Anthracnose Fungus Colletotrichum truncatum. Frontiers in Microbiology, 9(), 2367–. doi:10.3389/fmicb.2018.02367
- Rawlings, D., Barrett, J., Thomas, D., Huang, X., Bateman, A. & Finn, R.D. (2018) The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. Nucleic Acids Res 46, D624-D632.
- Rawlings, D., Waller, M., Barrett, J., and Bateman, A. (2014). MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Res. 42, D503–D509. doi: 10.1093/nar/gkt953
- Rebollo, R., Romanish, M., Mager, D. (2012). Transposable Elements: An Abundant and Natural Source of Regulatory Sequences for Host Genes. Annual Review of Genetics, 46(1), 21–42. doi:10.1146/annurev-genet-110711-155621
- Reuter, J., Spacek, D,. Snyder, M. (2015). High-Throughput Sequencing Technologies. Molecular Cell 58. P:586-597
- Rivano, F., Martinez, M., Cevallos, V., & Cilas, C. (2010). Assessing resistance of rubber tree clones to Microcyclus ulei in large-scale clone trials in Ecuador: A less time-consuming field method. European Journal of Plant Pathology, 126(4), 541–552.
- Rivano, F., Mattos, C., Cardoso, S., Martinez, M., Cevallos, V., Le Guen, V., & Garcia, D. (2013). Breeding Hevea brasiliensis for yield, growth and SALB resistance for high disease environments. Industrial Crops and Products, 44, 659–670. doi:10.1016/j.indcrop.2012.09.005
- Rivano, F., Vera, J., Cevallos, V., Almeida, D., Maldonado, L., & Flori, A. (2016). Performance of 10 *Hevea brasiliensis* clones in Ecuador, Under South American Leaf Blight escape conditions. Industrial Crops and Products, 94, 762–773
- Rocha, H.M.; Medeiros, A.G. Vasconcelos, A.P. 1978. Comparação de fungicidas para controle do mal-dos-folhas de seringueira (Microcyclus ulei (P. Henn.) v. Arx) em viveiro. Fitopatologia Brasileira. 3: 163.

- Rodriguez, L., Ebert, M., Bolton, D., Thomma B. (2018). Tools of the crook-infection strategies of fungal plant pathogens. Plant J 93:664–674.
- Ruan, J., & Li, H. (2019) Fast and accurate long-read assembly with wtdbg2. Nat Methods doi:10.1038/s41592-019-0669-3
- Ryder, L and Talbot, N. (2015) Regulation of appressorium development in pathogenic fungi. Current Opinion in Plant Biology, 26:8–13
- Sandhya, V., Rajesh, K., Gazara, Shadab, N., Sabiha, P., Debasis, C., Praveen, K.V (2016). Draft genome sequencing and secretome analysis of fungal phytopathogen Ascochyta rabiei provides insight into the necrotrophic effector repertoire Nature scientific reports. 6:24638 ;1-7
- Sato, M., Ogura, Y., Nakamura, K., Nishida, R., Gotoh, Y., Hayashi, M., Hisatsune, J., Sugai, M., Takehiko, I., Hayashi, T. (2019). Comparison of the sequencing bias of currently available library preparation kits for Illumina sequencing of bacterial genomes and metagenomes. DNA Research, (), dsz017–. doi:10.1093/dnares/dsz017
- Scharf, D., Heinekamp, T., Brakhage, A., Heitman, J. (2014). Human and Plant Fungal Pathogens: The Role of Secondary Metabolites. PLoS Pathogens, 10(1), e1003859–. doi:10.1371/journal.ppat.1003859
- Schubert, K., Ritschel, A. & Braun, U. (2003). A monograph of *Fusicladium* s. lat. (Hyphomycetes). Schlechtendalia, 9, 1-132.
- Schultzhaus, Z., S., & Shaw, D. (2015). Endocytosis and exocytosis in hyphal growth. Fungal Biology Reviews, 29(2), 43-53. https://doi.org/10.1016/j.fbr.2015.04.002
- Schuster, C. (2007). Next-generation sequencing transforms today's biology. Nat. Methods 5, 16–18
- Schuster, S.C. (2007). Next-generation sequencing transforms today's biology. Nat. Methods 5, 16–18
- Sedlackova, T., Repiska, G., Celec, P. et al. (2013). Fragmentation of DNA affects the accuracy of the DNA quantitation by the commonly used methods. Biol Proced Online 15, 5 . https://doi.org/10.1186/1480-9222-15-5
- Shirasu, A (2015). Plant cells under siege: plant immune system versus pathogen effectors. Current Opinion in Plant Biology 28: 1–8.
- Sibley, R., Blazquez, L., & Ule, J. (2016). Lessons from non-canonical splicing. Nature Reviews Genetics*, 17(7), 407–421.* doi:10.1038/nrg.2016.46
- Silva, K., S. Deraniyagala, A., Wijesundera, C., Karunanayake, Priyanka, S.(2002). Isolation of scopoletin from leaves of Hevea brasiliensis and the effect of scopoletin on pathogens ofH. brasiliensis. , 153(4), 199–202. doi:10.1023/a:101491013259
- Silva, M., Barreto, W ., Pereira , L., Freitas, M., Groenewald, Z., Crous, P. (2016). Exploring fungal mega-diversity: Pseudocercospora from Brazil. Persoonia 37: 142–172
- Simão, F., Waterhouse, R., Evgenia, P., Kriventseva, E., Zdobnov, M. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, Bioinformatics, Volume 31, Issue 19, Pages 3210–3212
- Sivashankari, S., & Shanmughavel P. (2006). Functional annotation of hypothetical proteins - A review. Bioinformation. 2006 Dec 29;1(8):335-8. doi: 10.6026/97320630001335. PMID: 17597916; PMCID: PMC1891709.
- Skamnioti, F., & Sj, G. (2008). Evolutionary history of the ancient cutinase family in five filamentous Ascomycetes reveals differential gene duplications and losses and in Magnaporthe grisea shows evidence of sub- and neo-functionalization. The New Phytologist; New Phytol. https://doi.org/10.1111/j.1469-8137.2008.02598.x

References

- Smit, A., Hubley, R., Green, P. (2015). Repeat masker Open-4.0 . Available from: http://repeatmasker.org
- Soanes, D. M., Alam, I., Cornell, M., Wong, H. M., Hedeler, C., Paton, N. W., … Talbot, N. J. (2008). Comparative Genome Analysis of Filamentous Fungi Reveals Gene Family Expansions Associated with Fungal Pathogenesis. PLoS ONE, 3(6), e2300. doi:10.1371/journal.pone.0002300
- Sonah, H., Deshmukh, R. K., & Bélanger, R. (2016). Computational Prediction of Effector Proteins in Fungi: Opportunities and Challenges. Frontiers in Plant Science, 7(e3), 390–14.
- Spanu, D., Abbott, C., Amselem, J., Burgis, A., Soanes, M., Stüber, K., Ver Loren, van Themaat, E., Brown., K., Butcher, A., Gurr, J., et al. (2010). Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism. Science. 2010 Dec 10;330(6010):1543-6. doi: 10.1126/science.1194573. PMID: 21148392.
- Sperschneider, J., Dodds, P. N., Singh, K. B., Taylor, J.(2018). ApoplastP: prediction of effectors and plant proteins in the apoplast using machine learning. New Phytol, 217(4): Pages 1764-1778.
- Sperschneider. J., Gardiner, D., Dodds, P., Tini, F., Covarelli, L., Singh, K., et al. (2015). EffectorP: predicting fungal effector proteins from secretomes using machine learning. New Phytol. 210(2): Pages 743–61.
- Stahl, E., Bishop, J. (2000) Plant–pathogen arms races at the molecular level. Current Opinion in Plant Biology 2000, 3:299–304
- Stajich, Jason E. (2017). Fungal Genomes and Insights into the Evolution of the Kingdom. Microbiology Spectrum, 5(4), –. doi:10.1128/microbiolspec.FUNK-0055-2016
- Stam, R., Jupe, J., Howden, A., Morris, J., Boevink, P., Hedley, P., Huitema, E., Arnold, D. (2013). Identification and Characterisation CRN Effectors in Phytophthora capsici Shows Modularity and Functional Diversity. PLoS ONE, 8(3), e59517–. doi:10.1371/journal.pone.0059517
- Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res; (Web Server issue): Pages 4–9.
- Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. 2006. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res; 34(Web Server issue):W435–9. Epub 2006/07/18. https://doi.org/10.1093/nar/gkl200 PMID: 16845043; PubMed Central PMCID: PMCPmc1538822.
- Staples, R. (2001). Nutrients for a rust fungus: the role of haustoria. , 6(11), 0–498. doi:10.1016/s1360-1385(01)02126-4
- Stergiopoulos, I., & De Wit, P.(2009). Fungal Effector Proteins. Annual Review of Phytopathology, 47, 233-263.
- Stergiopoulos, I., & Wit, P. J. G. M. de. (2009, agosto 3). Fungal Effector Proteins.; Annual Reviews. https://doi.org/10.1146/annurev.phyto.112408.132637
- Sterling, A., & Rodríguez, C. (2018). Estrategias de manejo para las principales enfermedades y plagas del cultivo del caucho con énfasis en la amazonia colombiana. Bogotá: Instituto Amazónico de Investigaciones Científicas SINCHI.
- Sterling, A., Galindo, L., Suárez, D., Velasco, G., Andrade, T., & Gómez, A. (2019). Early assessing performance and resistance of Colombian rubber tree genotypes under high south American leaf blight pressure in Amazon. Industrial Crops and Products, 141, 111775.
- Sterling, A & Melgarejo, L. (2018). Leaf gas exchange and chlorophyll a fluorescence in *Hevea brasiliensis* in response to Pseudocercospora ulei infection.

Physiological and Molecular Plant Pathology, 103(), 143–150. doi:10.1016/j.pmpp.2018.07.006

- Stirling, D. (2004) DNA extraction from fungi, Yeast and Bacteria. In J. M. Barlert & D. Stirling (Eds) PCR protocols Method in Molecular Biology pp. 53-54
- Stukenbrock E, Croll D. (2014). The evolving fungal genome. fungal biology reviews journal 28: 1-12
- Stukenbrock E., Croll, D. (2014). The evolving fungal genome. Fungal biology reviews 28 (2014) 1e12
- Szabo, J., & Bushnell, W. (2001). Hidden robbers: The role of fungal haustoria in parasitism of plants. Proceedings of the National Academy of Sciences, 98(14), 7654–7655. doi:10.1073/pnas.151262398
- Talbot, N.J., Ebbole, D.J. and Hamer, J.E. (1993) Identification and characterization of MPG1, a gene involved in pathogenicity from the rice blast fungus Magnaporthe grisea. Plant Cell, 5, 1575 – 1590.
- Talbot NJ (2003) On the trail of a cereal killer: exploring the biology of Magnaporthe grisea. Annu Rev Microbiol, 57:177-202.
- Tanaka, S., Han, X and Kahmann. (2016). Microbial effectors target multiple step in the salicylic acid production and signaling pathway.frontiers in Plant Science (6)349
- Tang, C., Yang, M., Fang, Y., Luo, Y., Gao, S., Xiao, X., et al. (2016). The rubber tree genome reveals new insights into rubber production and species adaptation. Nature Plants, 2(6), 16073–10.
- TerBush, R., & Novick P. (1995). Sec6, Sec8, and Sec15 are components of a multisubunit complex which localizes to small bud tips in Saccharomyces cerevisiae. J. Cell Biol. 130, 299-312. 10.1083/jcb.130.2.299
- TerBush R., Maurice, T., Roth, D., & Novick P. (1996). The Exocyst is a multiprotein complex required for exocytosis in Saccharomyces cerevisiae. EMBO J. 15, 6483-6494.
- Tian, L., Sun, W., Li, J., Chen, J., Dai, X., Qiu, N., & Zhang, D. (2021). Unconventionally Secreted Manganese Superoxide Dismutase VdSOD3 Is Required for the Virulence of Verticillium dahliae. Agronomy, 11(1), 13. https://doi.org/10.3390/agronomy11010013
- Trapnell, C., Williams, B., Pertea, G., Mortazavi, A., Kwan,G. et al . Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms. Nat Biotechnol. 2010 May ; 28(5): 511–515
- Tudzynski, P & Sharon, A. (2003). Fungal Pathogenicity Genes. Applied Mycology and Biotechnology (Vol. 3, pp. 187–212). Elsevier.
- Urban, M., Cuzick, A., Rutherford, K., Irvine, A., Helder, P., Pant, Ra., Sadanadan, V, Khamari, L., et al. (2017) PHI-base: a new interface and further additions for the multi-species pathogen–host interactions database, Nucleic Acids Research, Volume 45, Issue D1, January 2017, Pages D604–D610, https://doi.org/10.1093/nar/gkw1089
- Van Bel, M., Proost, S., Van Neste, C. et al. TRAPID: an efficient online tool for the functional and comparative analysis of de novoRNA-Seq transcriptomes. Genome Biol 14, R134 (2013). https://doi.org/10.1186/gb-2013-14-12-r134
- Venkatachalam, P., Geetha, N., Sangeetha, P., Thulaseedharan, A. (2013) Natural rubber producing plants: An overview. African J Biotechnol; 12: 1297–1310. https://doi.org/10.5897/AJBX12.016
- Viklund H, Elofsson A. Best alpha-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information.

References

Protein Sci. 2004;13(7):1908-1917. doi:10.1110/ps.04625404

- Vogel, J. (2008). Unique aspects of the grass cell wall. Current Opinion in Plant Biology; Curr Opin Plant Biol. https://doi.org/10.1016/j.pbi.2008.03.002
- Vurture, W. et al. GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics 33, 2202–2204 (2017).
- Walker, J., Abeel, T., Shea, T., Priest, M., Abouelleil, A., Sakthikumar, S., et al. (2014). Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome
- Wang, H., Fewer, P., Holm, L., Rouhiainen, L., Sivonen, K. (2014). Atlas of nonribosomal peptide and polyketide biosynthetic pathways reveals common occurrence of non-modular enzymes. Proceedings of the National Academy of Sciences, 111(25), 9259–9264. doi:10.1073/pnas.1401734111
- Wang, Z., Lienemann, M., Qiau, M., Linder, M. (2010). Mechanisms of Protein Adhesion on Surface Films of Hydrophobin. Langmuir, 26(11), 8491–8496. doi:10.1021/la101240e
- Warren, L., Yang, C., Vandervalk, P. et al. (2015). LINKS: Scalable, alignment-free scaffolding of draft genomes with long reads. GigaSci 4, 35 . https://doi.org/10.1186/s13742-015-0076-3
- Warris, A., & Ballou, E. R. (2019). Oxidative responses and fungal infection biology. Seminars in Cell & Developmental Biology, 89, 34-46. https://doi.org/10.1016/j.semcdb.2018.03.004
- Waterhouse, A., Procter, J., Martin, D., Clamp, M., Barton, G. (2009). Jalview Version 2-a multiple sequence alignment editor and analysis workbench. Bioinformatics 25: Pages 1189-1191.
- Webster, J., & Weber, R., (2007). Introduction to fungí. Cambridge – United Kingdom.http://www.dbbe.fcen.uba.ar/contenido/objetos/webster 30521807395_1400021643840195.pd
- Wenger, 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. doi:10.1038/s41587-019-0217-9
- White, T., Bruns, T., Lee, S. & Taylor, J. (1990). Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. PCR protocols a guide to methods and applications Academic Press.
- Wicker, T., Oberhaensli, S., Parlange, F., Buchmann, J., Shatalina, M., Roffler, S., Ben-David, R., Doležel, J., Šimková, H., Schulze-Lefert, P., Spanu, P., Bruggmann, R., Amselem, J., Quesneville, H., et al (2013). The wheat powdery mildew genome shows the unique evolution of an obligate biotroph. Nature Genetics, 45(9), 1092–1096. doi:10.1038/ng.2704
- Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O. (2007). A unified classification system for eukaryotic transposable elements. , 8(12), 973–982. doi:10.1038/nrg2165
- Wit, P. J. G. M. de, Burgt, A. van der, Ökmen, B., Stergiopoulos, I., Abd-Elsalam, K. A., Aerts, A. L., Bahkali, A. H., Beenen, H. G., Chettri, P., Cox, M. P., Datema, E., Vries, R. P. de, Dhillon, B., Ganley, A. R., Griffiths, S. A., Guo, Y., Hamelin, R. C., Henrissat, B., Kabir, M. S., … Bradshaw, R. E. (2012). The Genomes of the Fungal Plant Pathogens Cladosporium fulvum and Dothistroma septosporum Reveal Adaptation to Different Hosts and Lifestyles But Also Signatures of Common Ancestry. PLOS Genetics, 8(11), e1003088. https://doi.org/10.1371/journal.pgen.1003088

- Xu, Z., Wang, H. (2007). LTR-FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. 2007;35
- Yamashita, S., Takahashi, S. (2020). Molecular Mechanisms of Natural Rubber Biosynthesis. Annu Rev Biochem, 2020 89:1, 821-851
- Yandell, M., & Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. Nature Reviews Genetics, 13(5), 329–342. doi:10.1038/nrg3174
- Yeadon, P. J., and D. E. Catcheside. 1995. Guest: a 98 bp inverted repeat transposable element in Neurospora crassa. Mol. Gen. Genet. 247:105–109.
- Yin, Y., Mao, X., Yang, J., Chen, X., Mao, F., Xu, Y. (2012). dbCAN: a web resource for automated carbohydrate- active enzyme annotation. Nucleic Acids Res. 40(Web Server issue): Pages 445–W51.
- Zaccaron, A. Z., & Stergiopoulos, I. (2020). First Draft Genome Resource for the Tomato Black Leaf Mold Pathogen Pseudocercospora fuligena. Molecular Plant-Microbe Interactions. https://doi.org/10.1094/MPMI-06-20-0139-A
- Zhang, C., Maryam R., Erfan S., Siavash, M. (2018). "ASTRAL-III: Polynomial Time Species Tree Reconstruction from Partially Resolved Gene Trees." BMC Bioinformatics 19 (S6): Page 153.
- Zhang, D., He, J., Haddadi, P., Zhu, j., Yang, Z., Ma, L. (2018). Genome sequence of the potato pathogenic fungus Alternaria solani HWC-168 reveals clues for its conidiation and virulence. BMC Microbiology, 18(1), 176–. doi:10.1186/s12866-018-1324-3
- Zhang, G., Zhang, Q., Zhou, D., Short, L., Wang G., Ma J., Li, F., Kong, T., Wang Z., et al (2017). A Verticillium dahliae Extracellular Cutinase Modulates Plant Immune Responses. Molecular Plant-Microbe Interactions. https://doi.org/10.1094/MPMI-06-17-0136-R
- Zhang, H., Jain, C., & Aluru, S. (2020). A comprehensive evaluation of long read error correction methods. BMC Genomics, 21(S6). doi:10.1186/s12864-020-07227-0
- Zhang, H., Yohe, T., Huang, L., Entwistle, S., Wu, P., Yang, Z., Busk, P., Xu, Y., Yin, Y. (2018). dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. Nucleic Acids Research, –. doi:10.1093/nar/gky418
- Zhang, N., Luo, J., Bhattacharya, D. (2017).Advances in Fungal Phylogenomics and Their Impact on Fungal Systematics. Advances in Genetics, Volume 100. ISSN 0065-2660
- Zhang, Q., Arbuckle, J., and Wessler, S.(2000). Recently extensive and preferential insertion of members of the miniature inverted-repeat transposable element family Heartbreaker (Hbr) into genic regions of maize. Proc. Natl. Acad. Sci. USA 97:1160 – 1165.
- Zhao, C., Waalwijk, C., de Wit, P. J. G. M., Tang, D., & van der Lee, T. (2013). RNA-Seq analysis reveals new gene models and alternative splicing in the fungal pathogen Fusarium graminearum. BMC Genomics, 14(1), 21. doi:10.1186/1471-2164-14-21
- Zhao, Z., Liu, H., Wang, C., & Xu, J.-R. (2013). Comparative analysis of fungal genomes reveals different plant cell wall degrading capacity in fungi. BMC Genomics, 14(1), 1-15. https://doi.org/10.1186/1471-2164-14-274
- Zhong, Z., Marcel, TC., Hartmann. E., Ma, X., Plissonneau, C., Zala, M., Ducasse, A., Confais, J., Compain, J., Lapalu, N. et al. (2017) A small secreted protein in *Zymoseptoria tritici* is responsible for avirulence on wheat cultivars carrying the Stb6 resistance gene.

References

- Žifčáková, L., & Baldrian, P. (2012). Fungal polysaccharide monooxygenases: new players in the decomposition of cellulose. Fungal Ecology, 5(5), 481–489. doi:10.1016/j.funeco.2012.05.001
- Zimin, V., Marçais, G., Puiu, D., Roberts, M., Salzberg SL., Yorke, A. (2013). The MaSuRCA genome assembler. Bioinformatics;29:2669–77.
- Zipfel, C. (2008). Pattern-recognition receptors in plant innate immunity. Current Opinion in Immunology, 20, 10-16.
- Zipfel, C. & Felix, G. (2005). Plants and animals: a different taste for microbes? Current Opinion in Plant Biology, 8, 353-360.