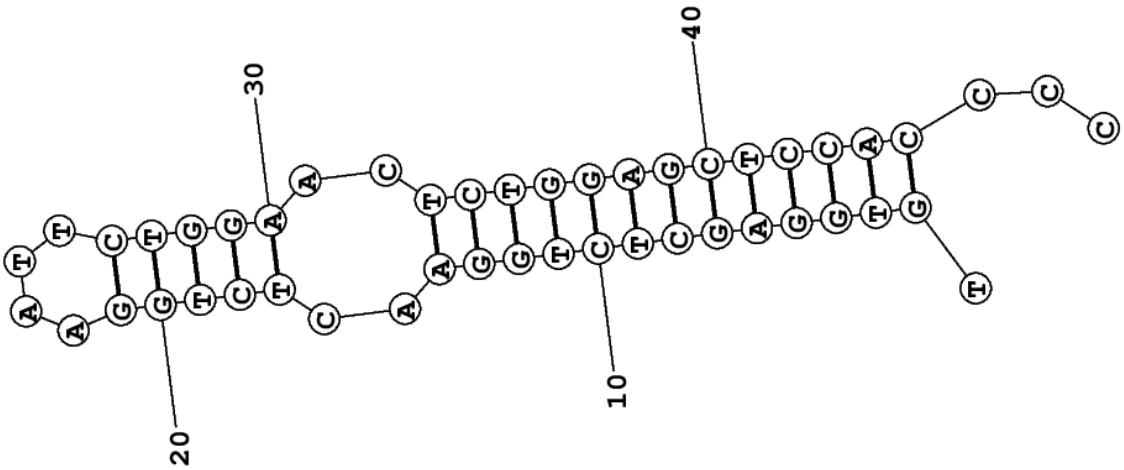
UNIVERSITAT ROVIRA i VIRGILI

PREDICTION AND CHARACTERIZATION OF MICRORNA-LIKE RNAS IN FUNGAL AND FUNGAL-LIKE PATHOGENS

Ivan Cardona Ferrer

FINAL DEGREE PROJECT

Double degree in Computer Engineering and in Biotechnology



Academic tutor: Javier Capilla Luque

In cooperation with: Biotech Vana, S.L.

Setembre 2024

Jo, Ivan Cardona Ferrer , amb DNI 20236883B, soc coneixedor de la guia de prevenció del plagi a la URV *Prevenció, detecció i tractament del plagi en la docència: guia per estudiants* (aprovada el juliol 2017) (http://www.urv.cat/ca/vidacampus/serveis/crai/que-us-oferim/formaciocompetencies-nuclears/plagi/) i afirmo que aquest TFG no constitueixen cap de les conductes considerades com a plagi per la URV.

Tarragona, 02 de setembre de 2024

IVAN| CARDONA| FERRER

Firmado digitalmente por IVAN|CARDONA| FERRER
Fecha: 2024.09.02 00:33:10 +02'00'

Index

Table Index

Figure Index

1 Info about the Center

Biotech Vana SL (or Biotechvana) [1] is a technology company based mainly on Bioinformatics and Computational Biology, although it implements other e-business models and ICT[1] services. It was created in 2006 by genetic researchers from the University of Valencia Carlos Llorens and Andrés Moya, together with the programmer Ricardo Futami and several collaborators and private capital.

Based on its main activity, the entity offers software products, omics analysis services and R+D+i consultancy to universities, research centers and hospitals, as well as to the biotechnology and pharmaceutical industry. Biotechvana's R+D+i efforts have been dedicated to developing its own software assets (source code and patents) and technical know-how protocols to position itself in the bioinformatics career. Among other assets, they have developed a software package that works both: a) linked to a server managing pipelines based on both free and proprietary software or b) as a standalone desktop software. In the same way, computational services are performed, including analysis and creation of database-type infrastructures.

In terms of organizational and financial matters, Biotech Vana SL is a micro-enterprise (SME[2]) located in the Scientific Park of the University of Valencia. The company is currently made up of a team of 7 permanent workers, with Carlos Llorens as CEO. In 2023, it obtained a turnover of around €190,000 and income from R&D projects worth €250,000 and has a share capital of over €100,000. It participates and has participated in several regional, national and European funding programs, including the Regional Operational Program of the European Regional Development Fund (FEDER) of the Valencian Community or the NextGenerationEU funds. The company also receives the support of the Government of Spain through the Ministry of Science and Innovation and the Secretary of State for Digitalization and Artificial Intelligence. In addition, it has been rewarded as an Innovative SME.

[1] Information and Communication Technology

[2] Small and Medium-sized Enterprises

1.1 Working position in the company and Self-assessment

I have worked on this project individually as a bioinformatician and biological data analyst. I have always received help from my supervisor in the company, Beatriz Soriano, who is a PhD. I received the aid specially at the beginning where the procedures to follow were explained to me, as well as the operation of the different bioinformatics tools to be used.

With the completion of this project, I have been able to participate in a complete bioinformatic analysis workflow, gaining hands-on experience in data preprocessing, small RNA prediction, target prediction, data annotation and characterization, as well as the use of various bioinformatic tools and biological databases. This experience has provided me with a comprehensive understanding of the processes involved in analysing high-throughput sequencing data and the application of these analyses to a real biological study.

It is also worth highlighting the use of my computational skills to create code scripts at different stages of the study, which have allowed me to process the information more quickly and efficiently.

It should also be noted that this work has been conducted alongside another project of similar characteristics, providing me with a broader perspective and highlighting the various possibilities available for data processing and analysis.

To sum up, the obtained results have met expectations and have helped to delve deeper into the subject of study and in turn facilitate future research.

2 Summary and Keywords

Bioinformatic analyses are fundamental to process biological data. They have become essential in many scientific fields, acquiring special importance in the agricultural and medical sector. The project hereby presented seeks to deeply analyse small RNA sequencing data from samples of different pathogenic agents that infect plants; more concisely, two species of fungi (*Botrytis cinerea* and *Fusarium oxysporum*) and two species of oomycetes (*Phytopthora infestans* and *Phytopthora parasitica*) are studied. The selected data comes from different experiments selected in a random manner from the NCBI Sequence Read Archive, always choosing wild type samples and with no contamination if possible. The treatment of the data is carried out using bioinformatic methods such as machine-learning, along with several tools (e.g. data preprocessing programs, micro RNAs –miRNAs-- and microRNA-like RNAs predictors, secondary structure predictors, mappers, omic annotators) and some useful databases (e.g. NCBI[3], Ensembl, KEGG[4]). Among the main tasks realized in this work we identify microRNA-like RNAs, also known as milRNAs, in the studied species and predict its secondary structures, as well as potential gene targets and related data to perform characterization. It must be said that currently there is very little or no information about the presence and function of the sncRNAs[5] just discussed in fungal and fungal-like (oomycetes) pathogens. Therefore, this study could have a significant impact on the understanding of the role of milRNAs in fungal and oomycete pathogenesis. The results obtained in this project could help to identify in the future new biomarkers for diagnosing and monitoring related diseases caused by these pathogens and could lead to the development of new antifungal drugs and new therapies.

Keywords: microRNAs; microRNA-like RNAs; fungi; oomycetes; plant pathogens; characterization; prediction.

[3] National Center for Biotechnology Information

[4] Kyoto Encyclopedia of Genes and Genomes

[5] small non-coding RNAs

3 Introduction

Nowadays the agricultural sector suffers from significant losses due to a great variety of infectious agents, such as bacteria, virus, fungi and protists, that infect the plants. These agents cause early crop deterioration, leading to a decrease in the productive system.

Among them, fungal and fungal-like pathogens are of particular importance. They are widely extended and can attack a broad range of host vegetal species. Furthermore, these organisms have an efficient reproduction capacity through spores, which can be spread easily by wind, water, soil, and insects, allowing them to infect large areas quickly. The spores can survive in harsh environmental conditions for long periods of time, lying dormant in soil or plant debris until conditions are favourable for infection. In addition, these pathogens have the ability to go through multiple infection cycles in a single growing season, leading to repeated attacks and compounding damage.

All the aforementioned factors, alongside with their diverse infection strategies and the capacity of causing extensive damage and destruction in almost all plant tissues, makes of fungi and fungal-like pathogens (oomycetes) the most dangerous organisms for crops. It needs also to be said that fungal related diseases provoked by mycosis and oomycetomycosis can be challenging to manage because they often require a combination of different treatments. In some cases, the infectious agent develop resistance to fungicides, further complicating control efforts, and also having an impact on the human clinical research area [2].

Therefore, this represents a worldwide concern that demands for a solution. There have been several approaches and efforts to solve this problem over the years, in order to reduce the economic and social impact of the infections caused by the pathogenic agents mentioned before. One key point has been the study of the role of promoter genes involved in plant defence [3].

Another more recent approach has been the research on the playing role of miRNAs and milRNAs in the interaction between plants and pathogens during an infectious cycle [4].

Nevertheless, the study of this type of sncRNAs represents a real challenge when it comes to its presence in fungi and, specially, in oomycetes. There are very few species belonging to these group of organisms where canonical miRNAs have been identified. In the recent years, some milRNAs have been found in many fungi and oomycete species. These RNA related structures could have similar functions to miRNAs, but do not meet all the criteria to be classified as canonical miRNAs and, in those groups where their presence have been denotated, there is little information about them [5], [6].

It is known that well-defined miRNAs from other species (specially plants and mammals) regulate gene expression at a posttranscriptional level through complementary base pairing with target mRNAs, leading to mRNA degradation and therefore blocking translation. Thus, miRNAs represent an epigenetic mechanism that controls gene expression in various homeostatic processes and pathological conditions within cells. Furthermore, the dysfunction of miRNAs has been associated with a large number of diseases [7], [8], [9].

The study presented here aims to provide further insights into the playing role and potential involvement of milRNAs in the development of infections in crops caused by fungal and fungal-like pathogens.

3.1 Pathogenic Agents under Study

Here, we now present the infectious agents that will be studied in this project.

The selection criteria of the organisms to study consisted first of a choice between four of the most economically relevant crops worldwide, especially in Europe, that are infected mainly by fungal or fungal-like pathogens:

- Wine Grapes (*Vitis vinifera*). Wine grapes are economically significant due to the global wine industry. They have substantial cultural and economic impact, especially in regions known for wine production such as Spain, Italy and France, which export this product around the world.
- Tomatoes (*Solanum lycopersicum*). Tomatoes are a key ingredient in a wide range of cuisines globally. They have high economic value due to their use in fresh consumption and processed products (e.g., sauces).

- Potatoes (*Solanum tuberosum*). Potatoes serve as a fundamental food source for millions of people globally. They have a high production volume, significant nutritional value, and versatility in cooking and processing. They are a crucial food security crop in many regions.
- Citrus. Citrus fruits, including oranges, lemons, limes, and grapefruits, are important for their vitamin C content and other nutrients. They are economically valuable and widely consumed both fresh and as juice. Among this group of fruits, we have selected *Citrus clementina*.

After aiming the target crops, a search was carried out and a relevant pathogen of the desired type (fungal or fungal-like) was chosen for each one.

3.1.1 Botrytis cinerea

Botrytis cinerea is a necrotrophic fungus belonging to the *Ascomycete* phylum and is responsible for causing grey mold disease [10]. It is widespread in all continents and can infect a wide range of plants in any stage of growth. The most notable host is the wine grape where causes bunch rot, but *B. cinerea* can also lead to important economic losses in several crops, including vegetables (e.g. lettuce), fruit (e.g. berries) and ornamentals (e.g. rose).

This fungus attacks weak plants or dying flowers. In fact, it helps the recycling process of plants by promoting their decomposition and therefore the availability of their nutrients in the substrate, thus playing a vital role in the natural growth cycle. It can be beneficial (noble rot) for sweet wine production under certain conditions too [11]. However, when *B. cinerea* attacks crops, it becomes a pest.

To detect an organism infected by *Botrytis* we can take a look at the plant itself. The tissue in which it develops darkens and sometimes softens due to the death of host cells. Over time, a layer of hairy, grey mould will appear on these areas (see *Figure 1*).

The development of the infection usually starts in plant remains from previous crops that were already infected and were not eliminated at the beginning of the new harvest. The mycelium present in these remains begins to develop with the increment of temperatures (as usually happens in early spring). In broad daylight, the mycelium begins to produce structures called conidiophores (see *Figure 2*).

At the end of these structures, spores called conidia are formed, which are transported through the air and can come into contact with the leaves and stems of the crop, thus infecting the plant.

Figure 1. *Vitis vinifera* fruit (grapevine) infected by *Botrytis cinerea*.

Source: https://www.canna.es/articles/botrytis-cinerea-plagas-y-enfermedades

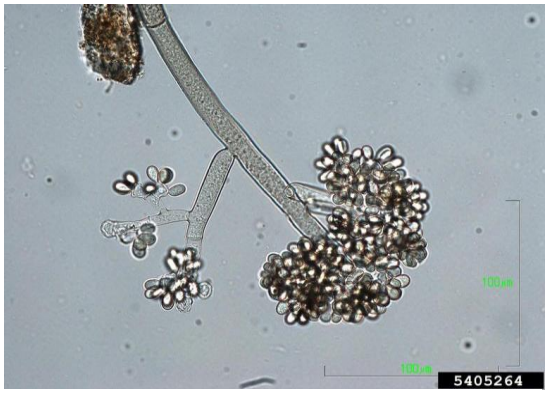


Figure 2. *Botrytis cinerea* conidiophores seen under a microscope.

Source: https://www.forestryimages.org/browse/detail.cfm?imgnum=5405264

3.1.2 Fusarium oxysporum

Fusarium oxysporum forms part of a group of particular fungi inhabiting the soil (see *Figure 3*). They can exist as saprophytes taking advantage of the soil debris but also as plant endophytes that colonize the plant roots.

This fungus has numerous different special forms or strains that are selectively pathogenic in a limited number of crops, being *Fusarium oxysporum f.sp. lycopersici* the most significant, which causes vascular wilt disease in tomatoes [12]. In the same crop, even different special forms may occur and show different symptoms, but all of them have vascular wilt in common. First, the leaves turn yellow and wilt, usually on one side of the plant, and eventually the entire plant wilts (see *Figure 4*). Other symptoms include brown discoloration of xylem tissues, which is seen when stems are cut.

F. oxysporum is the only *Fusarium* that actually grows within the vascular system of the host plant and spreads upwards within the plant, making it more dangerous. The other species spread upwards outside the plant.

Most *Fusarium* species only produce asexual spores, but there are some that also produce ascospores. The life cycle of *Fusarium oxysporum* is similar to that of most *Fusarium* species. It overwinters for several years in the soil and in infected plant debris, as chlamydospores (thick-walled mycelial cells) or mycelium. Survival is also possible on seeds, greenhouse structures, tools and machines. Primary infection is either seed-borne or occurs as a root infection at the root tip or in small wounds, for example where roots branch off the main root.

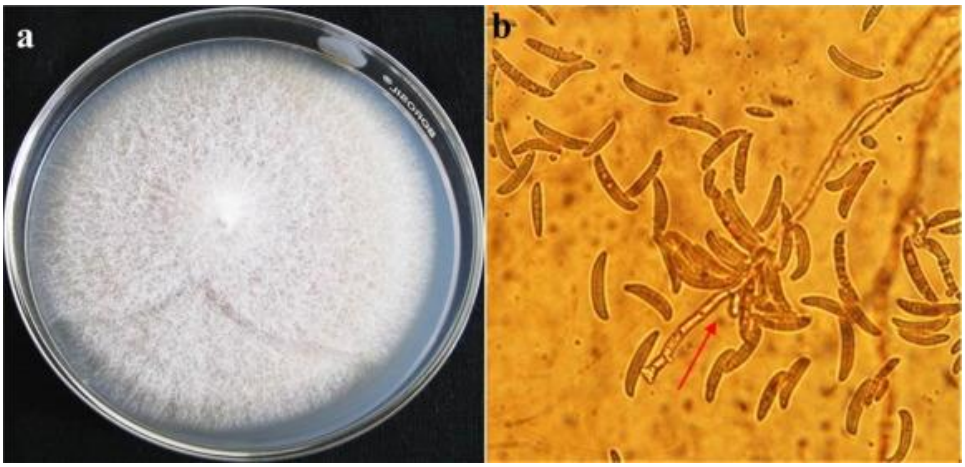


Figure 3. *Fusarium oxysporum* grown on a culture plate (a) and microscopical vision (b).

Source: https://www.sciencedirect.com/science/article/pii/S1319562X19301007

Figure 4. Vascular wilt caused by *Fusarium oxysporum*.

Source: https://www.koppert.es/enfermedades-de-las-plantas/marchitez-vascular/

3.1.3 Phytophthora infestans

Phytophthora infestans is a fungus-like oomycete (water mold) that causes late blight disease in potatoes and infects other members of the *Solanaceae* family, such as tomatoes [13].

The disease usually affects the leaves, stems and tubers. The initial sign of infection is green or grey spots that develop on the old lower leaves of the plant. As the disease progresses, the lesions turn either brown or black and spread to leaves and stem tips, where moisture is higher (see *Figure 5*).

This pathogen can affect large areas due to its ability to reproduce asexually through spores that are carried by the wind, causing major crop losses. In moist conditions, *Phytophthora infestans* produces sporophores enclosed in sporangium on the exterior side of affected parts. The sporophores develop into white spores that are noticed at the edges of the lower surface of leaves.

When the environment is wet and cool, the above mentioned structures release zoospores that can swim until reaching and infecting more parts of the host. If left untreated, the disease spreads to all parts of the plant, resulting in great damage (wilting leaves is a common symptom of highly infected plants). When conditions are warmer, they have the ability to form a germ tube in the sporangium that can penetrate the plant.

See *Figure 6* for more clearance about the different structures.

Figure 5. Late blight disease caused by *Phytophthora infestans*.

Source: https://cropaia.com/blog/potato-blight/

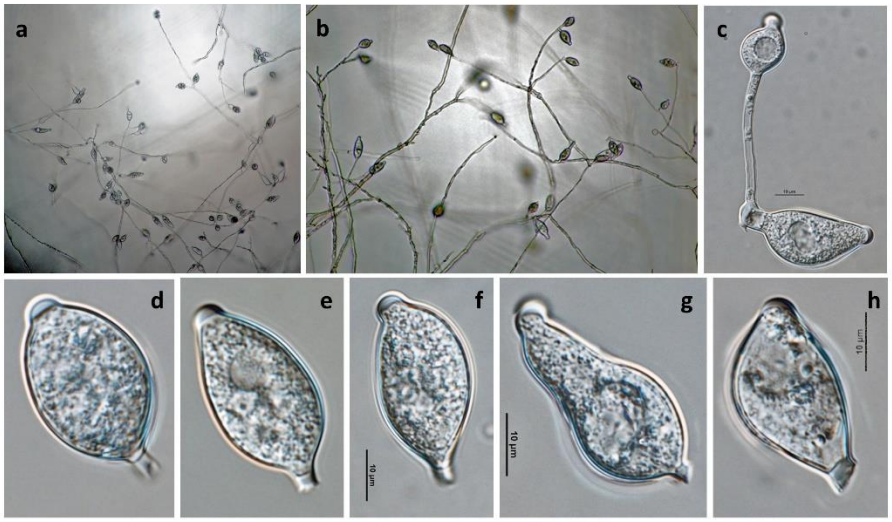


Figure 6. *Phytophthora infestans* asexual phase: (a, b) sporangia in sporangiophores; (c) sporangia external germination; (d–h) papillate sporangia with caducous short pedicels;

Source: https://idtools.org/phytophthora/index.cfm?packageID=1131&entityID=5087

3.1.4 Phytophthora parasitica

Phytophthora parasitica, currently known as *Phytophthora nicotianae var. parasitica*, is a highly destructive pathogen [14] that affects a wide variety of crop plants such as tomatoes, peppers, potatoes, tobacco, cacao, bananas, citrus, walnuts and almonds. This fungus-like oomycete also causes diseases in numerous nursery and ornamental plants, and has the capacity to injure entire forest ecosystems [15].

The infection begins with root rot, which retards growth and discolours leaves. In young plants it spreads to the neck, causing black neck in tobacco or neck rot in tomatoes, for example. In citrus it causes brown rot of fruit [16] with water splashes (concentric dark-colored buds --see *Figure 7*--) and, more commonly, gummosis [17] and rot in the lower part of the trunk (see *Figure 8*).

There are important structures this pathogen uses in its disease cycle. Chlamydospores are produced asexually and serve as long lived resting structures, surviving several years. In the same way as *P. infestans*, these spores germinate in warm and moist soil to produce a germ tube that infects plants or produces a sporangium, another asexual structure. Zoospores can be released (see *Figure 9*), which navigate through the water to the root tips. Once the root surface is contacted, zoospores encyst, and a germ tube will emerge penetrating the epidermis, provoking the start of the infection.

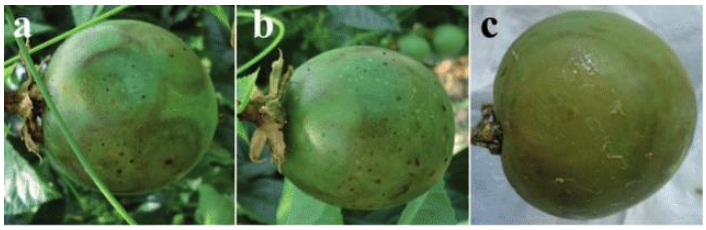


Figure 7. Brown rot of fruit caused by *Phtophthora parasitica*.
Source: https://www.researchgate.net/figure/Phytophthora-nicotianae-var-parasitica-Colletotrichum-gloeosporioides-sensu-lato-and_fig2_318117032

Figure 8. Typical symptoms caused by *Phytophthora parasitica*

Source:
https://www.researchgate.net/figure/Figura-34-Sintomas-tipicos-causados-por-Phytophthora-parasitica-Izq-y-P-cinnamoni_fig21_317356316

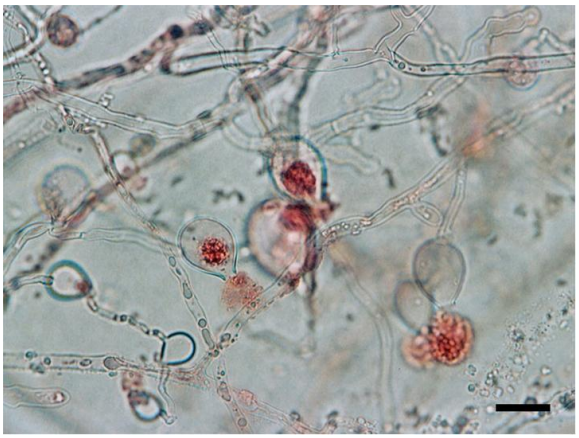


Figure 9. *Phytophthora parasitica* sporangia and zoospores.

Source:
https://commons.wikimedia.org/wiki/File:Phytophthora_parasitica_sporangia_and_zoospores.jpg

4 Working Hypothesis and Objectives

The working hypothesis of this project is that a complete bioinformatic analysis realized over different plant pathogens will lead to further knowledge about the presence of miRNAs and milRNAs in fungi and relatives, as well as the role that these potential regulatory biomarkers could play in infective processes.

Therefore, the main objective of the presented work is the identification and characterization of milRNAs in the studied species responsible for pathogenesis.

At the same time, this principal objective could be subdivided into more specific points or tasks addressed in the project:

1) Preprocessing of data by applicating bioinformatic methods to the sequencing data such as quality analysis and read preprocessing.
2) Mapping of data.
3) Identification of milRNA sequences.
4) Analysis of the secondary structures in the identified milRNAs.
5) Prediction of gene targets using complete databases.
6) Functional annotation of predicted targets.
7) Integration of related data in the obtained results.

Once we comprehend the working hypothesis and the objectives of the project, it could be said that high-quality preprocessing of samples enables effective mapping and prediction of the milRNAs in the studied pathogens. Furthermore, the analyses of secondary structures with optimal energy profiles, alongside target gene annotation and integration of Gene Ontology (GO) terms and metabolic pathways, will facilitate understanding of the possible functional roles of these small non-coding RNAs in pathogen-host interactions.

5 Methodology

In order to complete the different tasks involved in the analysis, we have made use of several bioinformatic tools and packages, as well as some biological databases.

Biotechvana's GPRO software suite plays a leading role in this work. GPRO Suite is a bioinformatic project developed and maintained by Biotechvana with the aim to provide Graphical User interface (GUI) customized solutions for omic data analysis in remote servers (the Cloud) or in the user PC. It includes six Client-Desktop applications and a server-side infrastructure with distinct dependencies such as databases, scripts, and Command Line Interface (CLI) third party software allowing different pipelines and workflow combinations based on the State-of-the-Art.

More specifically, throughout the project, we have utilized **RNASeq**[6] and **Worksheet**[7] programs from the commented software suite. RNASeq [18] is a desktop application to manage server-side pipelines and workflows for differential expression and enrichment analysis based on State-of-the-Art software under two modes of execution: pipeline-like and Step-by-Step-like. On the other hand, Worksheet is a desktop application for knowledge discovery and interrogation of omic annotations.

The raw data information used in this project comes from the Sequence Read Archive (SRA) of the NCBI. Specifically, four SRA of each studied pathogen species have been chosen. These SRA belong to different experiments selected in a random manner from the database, always choosing wild type samples and with no contamination if possible. All the procedures that will be presented below have been executed separately for each pathogenic organism.

[6] RNASeq used version: 2.3.2

[7] Worksheet used version: 2.3.1

5.1 Quality Analysis

In a bioinformatics project, quality analysis is a crucial step to ensure that the data obtained is reliable and useful for further analysis [19]. This phase aims to assess and improve the quality of the data before proceeding with more detailed protocols.

In the presented work, we have used a tool integrated into the RNASeq app called FastQC, which will be explained below, in order to perform the quality analysis successfully.

5.1.1 FastQC

FastQC [20] is a quality control tool for high throughput sequence data. It aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. This tool provides a modular set of analyses which can be used to check whether the data has any problems that should be taken into consideration before doing any further analysis.

In this study, FastQC has been used to obtain a quality report in HTML format for each fastq file downloaded from the SRA database. It provides a quick overview of the different parameters analysed, so that the researcher can see the areas where there may be problems. Furthermore, the information is summarised in graphs and tables to quickly assess the data.

For more clearance, the results of a processed fastq file can be seen on *Figure 10*. In the image, the different parameters analysed by FastQC program are shown on the left side. For the sake of simplicity, only the complete results of the two first parameters are displayed.

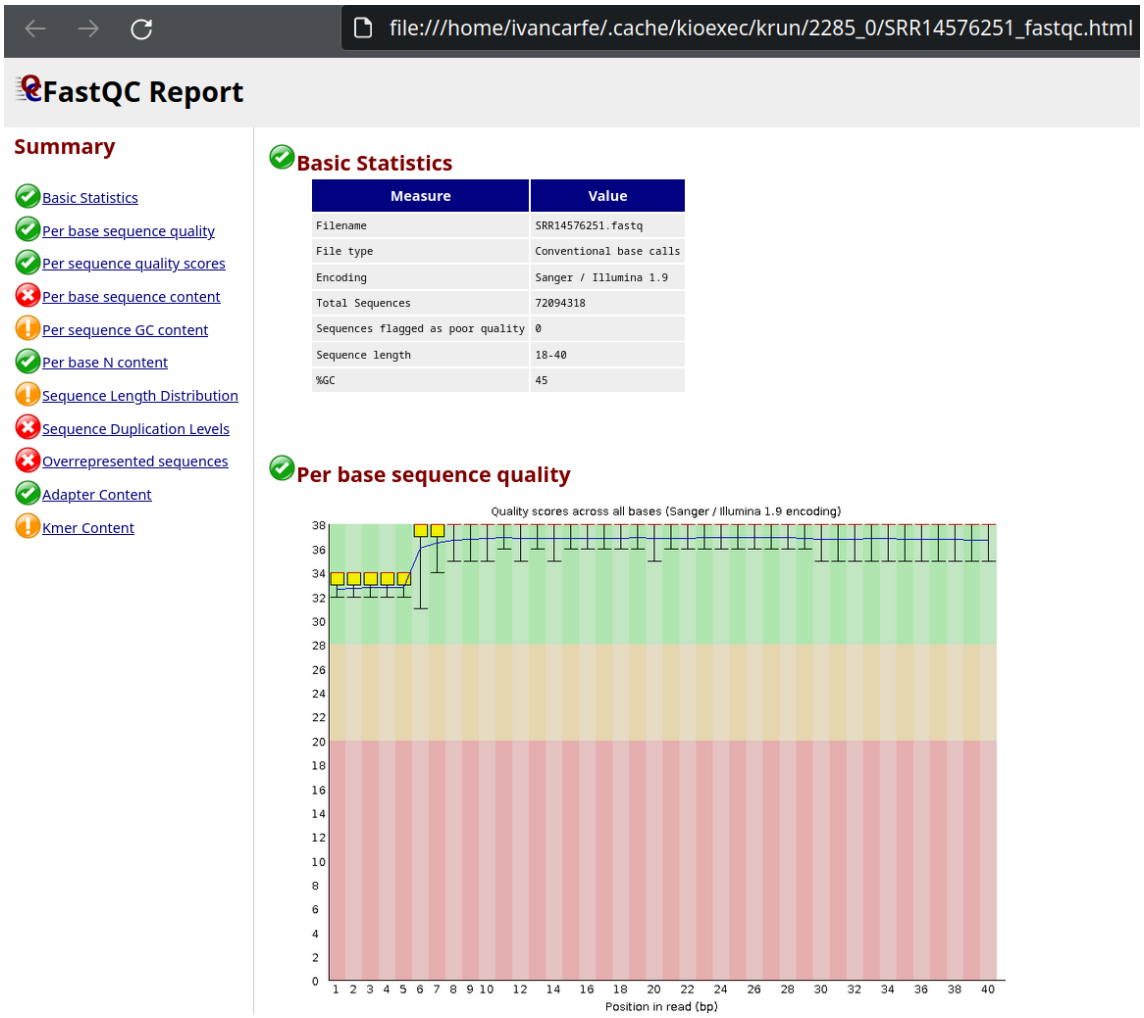


Figure 10. Capture of an html report generated by FastQC program after processing an input fastq file.

5.2 Preprocessing Reads

After performing a quality analysis, a significant amount of relevant information is generated, which can help the bioinformatic researcher in the precise treatment of the data to improve its quality.

In the studied cases presented in this work, some files showed the presence of overrepresented sequences, provoked mostly by adapters. These adapters can be trimmed from the sequences with an integrated RNASeq tool called Cutadapt.

5.2.1 Cutadapt

Cutadapt [21] finds and removes adapter sequences, primers, poly-A tails and other types of unwanted sequences from high-throughput sequencing reads.

Cleaning the data is often necessary because small-RNA sequencing reads can include 3' sequencing adapters, amplicon reads start with primer sequences, and poly-A tails, although useful for pulling out RNA, are usually unwanted in the reads. Cutadapt assists with these trimming tasks by locating adapter or primer sequences in an error-tolerant manner, modifying and filtering single-end and paired-end reads.

To execute this functionality, we first need to identify and retrieve the detected adapter sequences, including those part of overrepresented sequences, from specific databases (e.g. Illumina) and copy them into the "Adapters" field in the program. We opted to trim both the 5' and 3' ends to ensure thorough processing of the reads in the fastq files.

It is crucial to eliminate adapter content because it removes unwanted sequences that can interfere with read alignment, mapping, and downstream analysis, thereby improving the overall quality and reliability of the sequencing data.

After running Cutadapt, a quality analysis with FastQC was performed again to confirm the effective removal of the desired adapter sequences.

5.2.2 PRINSEQ

Upon removing the adapters, we proceeded with further cleaning and filtering with the PRINSEQ functionality integrated in the RNASeq application.

PRINSEQ [22] is a tool for quality control of genomic and metagenomic sequence data that contains functions for trimming, filtering, and data reformatting.

This represents the final step of the preprocessing phase. Either with the raw data or with the obtained fastq after being processed with Cutadapt, we applied a series of filters in order to improve the quality of the data for further analysis in the next phases:

- **min_len.** This represents the minimum sequence length that we want. After observing the FastQC results, we decided to set a value of 13, so sequences shorter than 13 bp were removed. By doing this, we make sure to leave out possible traces of contamination (minimum length of microRNAs is about 16, but we wanted to be a little more permissive).
- **min_qual_mean.** This refers to the minimum quality score mean, that is, the lowest average quality score observed across all base positions in the reads. We fixed it to 24, so those sequences with a lower average quality score were removed. This value represents a medium level of severity; there is approximately a 1 in 250 chance of an incorrect base call, which translates to about 99.6% accuracy.

We also applied some trimming options to eliminate ambiguous or unknown nucleotides (depicted as Ns in the sequences):

- **ns_max_p.** This refers to the maximum allowed Ns percentage. We established a value of 10, so those sequences with a higher Ns percentage were discarded.
- **trim_ns_left** & **trim_ns_right**. We set it to 1, trimming all contiguous Ns from both the 5' and 3' end of the sequences.

After running the job with the parameter configuration explained above, new fastq files were generated, supposedly with improved quality. We performed another FastQC analysis over the treated files to check this.

5.3 Genome Reference Indexing

Once the quality analysis has been performed and preprocessing methods have been applied over the studied sequences to improve their quality, new processing steps are required in order to obtain further information.

In most cases, there is a need to perform some mapping. "Mapping" in bioinformatics refers to the process of aligning sequence reads obtained from sequencing data against a known reference sequence, such as a genome or transcriptome. The goal of mapping is to determine the exact position where each read comes from within the reference sequence.

Nevertheless, performing mapping can be a costly process. This is why, on many occasions, genome indexing is conducted prior to mapping procedures. Indexing involves creating an optimized data structure that allows programs to quickly search for matches between read sequences and the reference genome. This process makes the search much more efficient, as it avoids the need to search through the entire genome sequentially, which would be much slower. In summary, the goal of indexing a genome is to improve the speed and efficiency of mapping reads against that genome.

In the present work, we have used a program called Bowtie to perform indexing, which will be explained below.

5.3.1 Bowtie

Bowtie[8] [23] is a software package commonly used for sequence alignment and sequence analysis in bioinformatics. It is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences.

With the aim of completing the objectives of this work, we want to generate a genome index for each studied pathogen with this tool. By launching a command using a specified genome sequence, Bowtie generates a series of files with a prefix and various extensions, including: ".1.ebwt", ".2.ebwt", ".3.ebwt", ".4.ebwt", ".rev.1.ebwt" and ".rev.2.ebwt".

These files contain different parts of the generated index that will be used lately to quickly map reads against the reference genome.

5.4 milRNAs Prediction

As said before, the main objective of this project is to identify potential milRNAs in the studied pathogenic species. Currently there are practically no micro-like RNAs described in these organisms, and therefore we need to make predictions.

With the indexes already generated from the previous step, we can map the processed fastq files against their respective genomes. These tasks haven been realized with a program called miRDeep2, which is explained below.

[8] Bowtie used version: 1.3.1

5.4.1 miRDeep2

miRDeep2[9] [24] is a software package for identification of novel and known miRNAs in deep sequencing data. It can be used for miRNA expression profiling across samples too. Mapping is performed with the `mapper.pl` script and miRNA identification is done thanks to the `miRDeep2.pl` script.

- **mapper.pl →**

This script takes the small RNA reads (from the selected pathogenic organisms in our case), filters them, collapses them to remove redundancies, and maps –aligns-- them against the reference genome (selected fungal pathogens) using the Bowtie tool. Mapping results are saved in an .arf file, which can be used in subsequent steps of the miRDeep2 pipeline to predict and quantify microRNAs.

Different parameters of the tool were configured to generate a file with the sequences not aligned to the reference genome, exclude duplicates, filter sequences with undefined nucleotides, and group all redundant reads into a single read. A minimum length of 18 nucleotides was set for the sequences to be mapped, and information about the original number of reads was preserved in the grouped sequences. In addition, previously generated indexes of the fungal genomes were used in each run to optimize the mapping.

- **miRDeep2.pl →**

This is the main command that runs the `miRDeep2.pl` script from the miRDeep2 package. This script is the primary tool for the prediction of miRNAs in small RNA sequences mapped to the reference genome. It takes the small RNA sequences mapped against the reference genome (e.g. *Botrytis cinerea*) and uses miRDeep2 to identify potential miRNAs. The script makes a prediction of the miRNAs present in the small RNA data, identifying both known miRNAs and potential new miRNAs based on mapping information and genome sequences. The results generate .bed files that indicate the position of the found microRNAs, and fasta files with the identified sequences. The execution of this command also creates a fasta intermediate file (`identified_precursors.fa`) that contains potential miRNAs precursors.

[9] miRDeep2 used version: 0.1.2

The precursors found with `miRDeep2.pl` command will be useful next to predict more microRNA-like RNAs. We can perform this prediction with the milRNApredictor program, explained below.

5.4.2 milRNApredictor

milRNApredictor[10] [25] is a bioinformatic tool used to predict milRNAs by analyzing precursor RNA sequences identified by other programs such as miRDeep2. The tool validates these sequences by comparing them to sets of positive and negative sequences to ensure that they have similar characteristics to known milRNAs. This allows researchers to identify and study potential milRNAs in fungi and derivatives.

In the present study, we have utilized the identified miRNA precursors from the previous miRDeep2 runs, along with a positive and a negative dataset obtained from different well-identified fungi milRNAs and pseudo-milRNAs, respectively.

The executing command is called `milRNApredictor.pl` and uses the milRNApredictor software to identify potential milRNAs. It gets the integrated positive and negative data sets of known sequences to train the prediction model and then applies this model to the specific precursor sequences to make the predictions, assigning to each sequence a probability of being a milRNA.

5.5 milRNAs selection

In order to perform further processing with the identified microRNA-like RNAs, we first must filtrate the obtained results from the predictions. We will do so with a pair of tools explained below in this section.

5.5.1 Worksheet

Worksheet [26] is an app of the GPRO suite package made to supervise the downstream steps for data integration, prioritisation and knowledge discovery of results from omic annotations. It is a dynamic grid of columns and rows controlled by menu allowing the user to easily manage one or more annotation sets.

[10] milRNApredictor used version: 1.0.0

The software is coupled with the GPRO server infrastructure to call several knowledge databases from where Worksheet extracts the annotations. It is also able to link fasta files to the annotation file opened by Worksheet, and to filter or prioritise sequences according to the information provided by the annotations.

In this work, we aim to select those predicted milRNAs that have a significant potential to be truly so. Therefore, we need to establish a threshold value to eliminate those predictions that are less reliable.

First, we need to open as a .csv the output archive generated previously by milRNApredictor, which contains the predicted milRNAs, in the Worksheet app. Those predicted milRNA closer to the negative dataset have a predict type of -1, and those closer to the positive dataset have a predict type of 1. We only want to select the positive predictions, but if we desire to be a little stricter, we can stablish a probability cutoff value greater than 0.5. Since we are dealing with hypothetical milRNAs, we have been quite lax and have turned up the filtering value to 0.55.

After that, we need to link the filtered predictions with the sequence database of the identified precursors. The file that acts as sequence database is the fasta with the identified precursors generated by the miRDeep2 program, which contains the sequences for all the milRNA predictions made with milRNApredictor. The linkage can be done through to the name associated to each milRNA.

Once the linkage was done, we exported the resulting sequences as a fasta file for further processing.

5.5.2 CD-Hit

Prediction programs often generate redundant data which must be eliminated to avoid processing the same information several times and reduce the running load. To solve this problem, in the present study we have made use of CD-Hit.

CD-Hit [27] is a software widely used in bioinformatics to cluster similar sequences and eliminate redundancy in sequence databases. It is primarily used to reduce the complexity of large data sets by identifying and grouping sequences that are identical or very similar, keeping a single representative sequence for each group.

Using this program (version 4.8.1), redundant sequences were removed from the milRNAs that passed the 0.55 filter using 100% similarity, thus leaving only one representative sequence for each set of identical sequences.

5.6 Secondary Structures Prediction

Once we have predicted milRNAs for the studied organisms and have eliminated the redundances, now we can move on to the prediction of their secondary structures with the aim of gaining more information.

We have used two different programs specialized in secondary structures prediction, so that the results could be compared later.

5.6.1 RNAstructure

RNAstructure[11] [28] is a complete package for RNA and DNA secondary structure prediction and analysis. It includes algorithms for the aforementioned purpose, including facility to predict base pairing probabilities, and uses a free energy model to fit the sequences. It can also predict the folding energy, find the most stable secondary structures, and analyse the probability of various structures.

In order to predict the secondary structures with this software we need to create a file called `Turbofold.config` which must contain the desired configurational parameters. TurboFold [29] is one of the programs within RNAstructure that is specifically used for the cooperative prediction of RNA secondary structures using multiple homologous sequences. This method integrates evolutionary and thermodynamic information to improve the accuracy of predictions.

The parameters used in this study for TurboFold analysis are as follows: The input RNA sequences are provided in a fasta format file, which TurboFold processes to make the structural predictions. The output consists of .ct files that store the predicted secondary structures for each sequence. The predictions are made at a temperature of 310.15 K (37 °C), with the analysis refined over three iterations, and a gamma parameter is used to adjust the energy considerations.

[11] RNAstructure used package version: 6.4

Furthermore, TurboFold utilizes multiple processors for parallel computing, and the MEA (Maximum Expected Accuracy) mode is employed for structure prediction, with a MeaGamma value of 1 for balanced accuracy. Additionally, the launched analysis limits the maximum percentage of sequences differing from the consensus to 50%, the maximum number of structures considered, and applies a specific window size to minimize noise in the results.

Once we had `Turbofold.config` file ready, we executed it as a script and then obtained the .ct files that contain the secondary structure predictions (linear description of base pairs). Due to the fact that we also wanted to obtain graphical representations of the results, we executed another command: `/opt/RNAstructure/exe/draw`. This command generates .ps files, which contain the predicted secondary structures represented in a graphical manner.

5.6.2 RNAfold

RNAfold [30] is a software from the ViennaRNA[12] package [31] that predicts RNA secondary structures using a minimum free energy (MFE) model. It also offers the possibility to calculate the probability distribution of paired bases in different possible structures (dot-bracket notation), as well as the prediction of pseudoknots in the secondary structures.

We have executed a RNAfold command with some configurational parameters to make the structural predictions: The program has been instructed to use the Maximum Expected Accuracy (MEA) method with a gamma value of 1, ensuring a balanced prediction approach. The input file is the same fasta file used with RNAstructure which contains the RNA sequences to be analyzed. The predicted secondary structures and relevant data are redirected to an output file.

Finally, the execution of the command generates .ps files of two types:

- ss.ps → These files contain a static graphical representation of the predicted RNA secondary structures.
- dp.ps → These files contain "dot plots" where base pairs are plotted in a matrix plot. Here, the coordinates of each dot indicate a possible base pair.

[12] ViennaRNA used package version: 2.6.4

5.7 Target Prediction

Target prediction is the process of determining which molecules are regulated by other molecules.

In the present study, we want to determine which messenger RNA (mRNA) molecules from the selected plants are regulated by the predicted micro-like RNAs of the pathogenic agents that infect them. This prediction will be useful for understanding better how milRNAs influence gene expression and thus the biological processes in which they are involved.

With the aim of finding targets in the chosen plant species infected by the selected fungi and oomycetes, we have used sRNAtoolbox software (explained below).

5.7.1 sRNAtoolbox (miRanda)

sRNAtoolbox[13] [32] is a collection of bioinformatic tools designed to analyse and process small RNA molecules (sRNAs), including microRNAs. This set of tools is widely used in the field of genomics to perform various tasks such as predictions, identification of targets, annotations, and functional analysis of sequences.

miRanda [33] is a bioinformatics tool belonging to the sRNAtoolbox package used to predict the interactions between miRNAs and their target sequences in the genes of an organism. The tool works by looking for regions of complementarity between miRNAs and mRNA sequences (target sequences). From these predictions, miRanda calculates a score and a binding energy that indicate the probability and strength of the interaction between a miRNA and its target.

In the present work, we have executed a command that uses the miRNAconsTargets tool of sRNAtoolbox to identify target sequences for the selected pathogenic organisms milRNAs in the chosen host plant species coding sequences. The miRanda algorithm is used to perform the predictions, and the results are saved lately in a specified directory. The CDS (coding sequences) archives of the plants utilized in this study can be downloaded directly from https://plants.ensembl.org/index.html .

[13] sRNAtoolbox used package version: 0.0.6

The execution carried out generates several files, but the only one that is of our interest is “miranda.txt”. This file contains the predictions results generated by miRanda tool. It stores detailed information about the predicted miRNA-mRNA interactions (identified microRNA, identified mRNA, interaction energy, target start and end position, miRanda score). An example can be seen in *Figure 11*.

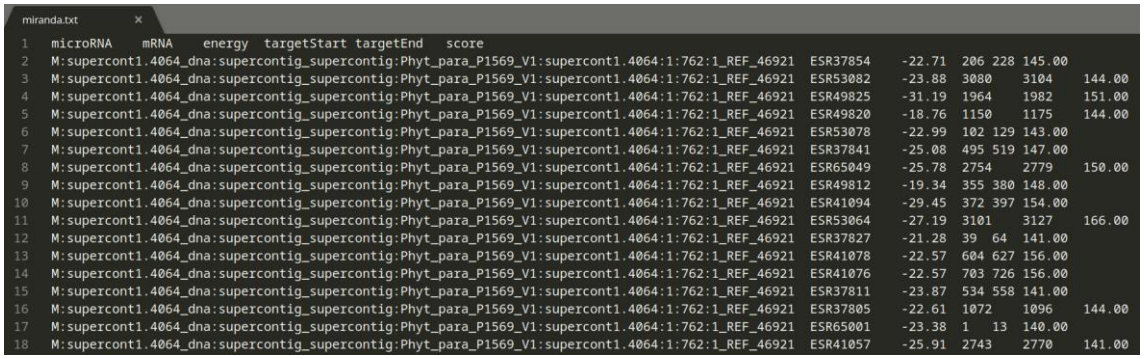


Figure 11. Screenshot of a miranda.txt file segment for *Phytopthora parasitica* target prediction.

Since we obtained files with a large amount of information, we decided to create a script to filter by energy and score values in a stricter manner. Initially minimum score was set to 140 and there were no minimum energy values established. That is why we raised the score value to 150 (the more score the more reliable the prediction) and the energy to -30 (the lower energy the more stable the union).

After that, we decided to code another script to select only the one prediction with the best score value among those targets that presented several affinity sites (same microRNA and mRNA but different target start and target end sites). This was done with the purpose of conserving only the most relevant information for further steps.

5.8 Functional Annotation

Now, we want to know which are the genes that encode the identified target mRNAs and gather related functional information to gain further insights.

We have used a platform called BioMart for the aforementioned purpose, which will be explained below.

5.8.1 BioMart

BioMart[14] [34] is a bioinformatics tool integrated in the EnsemblPlants web that provides an interface for biological data retrieval and filtering. It is a flexible and powerful platform that allows users to access large biomedical databases, such as genomes, proteomes, transcriptomes, and other types of omics data, easily and efficiently.

We want to use this functionality to annotate several informational parameters related to the previously identified targets. Therefore, the found targets were used to obtain data about their corresponding genes, including the gene name, description, type and associated GO terms.

Thus, we have not only predicted the targets for the milRNAs studied in this project, but also gathered valuable information about the genes encoding these plant transcripts, including their types and associated functions. This data is crucial as it identifies and characterizes the genes that produce the mRNAs interacting with the milRNAs of the pathogens.

Moreover, we also created GO annotations with the collected information about the genes. In these annotations we annotated each identified gene alongside with its description and type, followed by the GO accessions (with their terms and related functions). This GO annotation format is more compact, structured, and clear, showing all the GO terms associated to each single gene and therefore eliminating redundant data. It offers significant advantages in terms of readability, ease of analysis, and interpretation, especially when working with automated scripts or large datasets.

Furthermore, we integrated additional information such as the enzymatic codes associated to the GO accessions, as well as the related metabolic pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG). We were able to annotate all this complementary data thanks to the Worksheet app annotation functionalities.

[14] BioMart used version: Ensembl Genomes release 59 - May 2024

6 Results and Discussion

In this chapter, we will discuss the most relevant results obtained in the different phases of this bioinformatic analysis for each studied pathogenic species.

6.1 Quality Analysis Results

Before executing preprocessing, *Botrytis cinerea*, *Fusarium oxysporum* and *Phytophthora parasitica* presented good quality samples, with mean per base sequence quality values between 31 and 41 (values over 28 are considered as "good"). However, in the case of *Phytophthora infestans* the quality of the samples was lower, with mean per sequence quality values between 18 and 25; considered as medium quality values.

Regarding overrepresented sequences, all organisms but *Botrytis cinerea* presented significant adapter and/or primer content, which was effectively removed from the samples with Cutadapt.

After executing preprocessing with PRINSEQ, the three pathogens with good quality parameters didn't present remarkable improvements. Nevertheless, *Phytophthora infestans* results showed an actual improvement of the samples quality (see example in *Figure 12*).

In *Table 1*, the sequences that have been filtered after applying the preprocessing parameters explained in Chapter 5 are shown for each processed sample of the different organisms, corroborating the discussion of the results realized above.

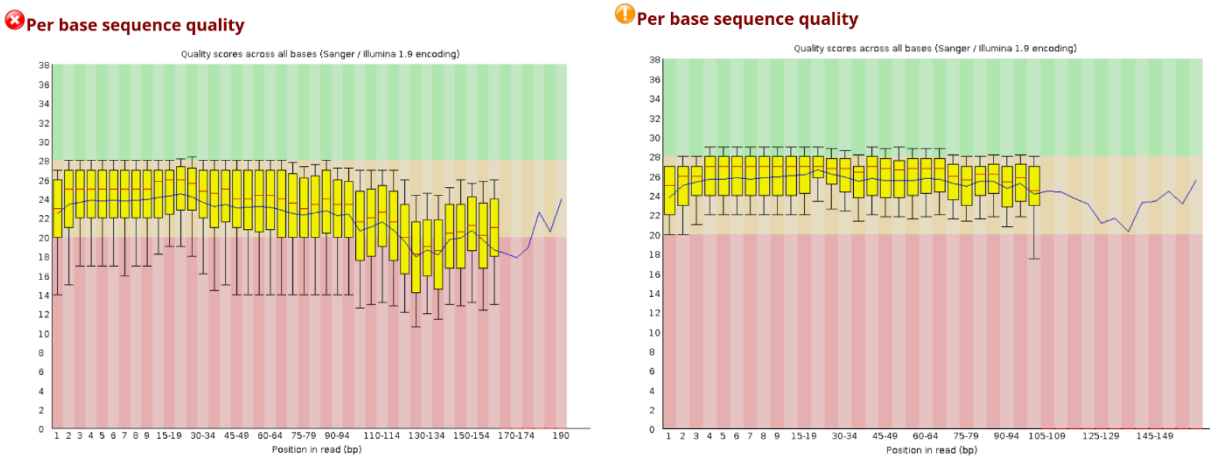


Figure 12. Per base sequence quality results before (a) and after (b) applying preprocessing techniques for sample SRR2819862 from *Phytophthora infestans*.

Table 1. Summary of sample results after applying preprocessing techniques with PRINSEQ.

***Botrytis cinerea* samples**	**Total sequences**	**% good sequences**	**Filtered sequences**
SRR14576251	72,094,318	99.36	459,239
SRR14576253	102,089,112	99.36	654,905
SRR24797115	517,747	99.93	358
SRR24797116	382,168	99.94	213
***Fusarium oxysporum* samples**	**Total sequences**	**% good sequences**	**Filtered sequences**
SRR27533538	47,548,893	99.27	346,080
SRR27533539	46,332,088	97.56	1,129,477
SRR27533540	47,081,719	99.56	209,124
SRR27533541	32,247,420	97.65	758,828
***Phytophthora infestans* samples**	**Total sequences**	**% good sequences**	**Filtered sequences**
SRR2819862	11,717,627	59.84	4,705,607
SRR2819864	12,872,161	67.60	4,170,002
SRR2819866	11,779,187	29.08	8,353,400
SRR2819871	8,812,459	64.44	3,133,511
***Phytophthora parasitica* samples**	**Total sequences**	**% good sequences**	**Filtered sequences**
SRR5437910	13,709,582	98.69	178,945
SRR5437911	19,258,731	98.31	325,090
SRR5437912	11,666,611	91.97	936,873
SRR14788321	17,086,946	100.00	150

All the HTML files generated by FastQC analyses can be found in the supplementary material (`01_quality_analysis` folders, before preprocessing; `02_preprocessed_reads` folders, after preprocessing).

6.2 Mapping Results

It is important to comment on the mapping that precedes both milRNAs predictions carried out because in this way we can obtain more information about the samples with respect to their reference genomes.

In *Table 2* the results obtained after executing the `mapper.pl` command from miRDeep2 for the different samples are showed. Complete detailed results can also be found in the supplementary material (`03_mapping` folders).

We can see that *Fusarium oxysporum* has the highest mapping percentages, indicating high data quality and good alignment with the reference genome. *Phytophthora infestans* and *Phytophthora parasitica* have lower mapping percentages, which could indicate greater alignment difficulty or greater genetic diversity within these species, or perhaps a less suitable genomic reference. In the case of *Botrytis cinerea*, the two first samples have a relatively high percentage of mapped reads, indicating good coverage and high compatibility between the reads and the reference genome. However, the other two samples show significantly lower percentages, which could indicate lower data quality, differences in experimental conditions, or potentially poor alignment with the selected reference genome.

Table 2. Mapping results obtained after executing mapper.pl command from miRDeep2.

***Botrytis cinerea* samples**	**Mapped reads**	**% mapped reads**
SRR14576251	49484332	69.080
SRR14576253	76709999	75.628
SRR24797115	92574	17.893
SRR24797116	49238	12.891
***Fusarium oxysporum* samples**	**Mapped reads**	**% mapped reads**
SRR27533538	38759634	82.851
SRR27533539	35202805	83.368
SRR27533540	38172498	81.964
SRR27533541	25593330	84.793
***Phytophthora infestans* samples**	**Mapped reads**	**% mapped reads**
SRR2819862	3067958	47.401
SRR2819864	3784229	46.660
SRR2819866	706999	35.857
SRR2819871	1808162	36.598
***Phytophthora parasitica* samples**	**Mapped reads**	**% mapped reads**
SRR5437910	4653916	34.739
SRR5437911	6050519	32.331
SRR5437912	2969014	32.505
SRR14788321	9621219	56.308

6.3 milRNAs Prediction Results

After executing the predictions of milRNAs with miRDeep2 and milRNApredictor, performing extraction/selection with Worksheet, and eliminating redundances with CD-Hit, we can analyse a series of results as seen in *Table 3* and compare the behaviour of both programs. Fasta files with the predicted milRNAs can be found in the supplementary material (`04_predicted_milRNAs` folders).

Table 3. Number of milRNAs predicted for the totality of the samples selected for each studied species using miRDeep2 and milRNApredictor software.

Pathogen	miRDeep2 (predicted milRNAs)	milRNApredictor (predicted milRNAs)
Botrytis cinerea	18	264
Fusarium oxysporum	11	530
Phytophthora infestans	21	184
Phytophthora parasitica	14	71

milRNApredictor is specifically designed to detect milRNAs in organisms where the known patterns of microRNAs may not be well established, such as fungi and derivatives. Therefore, this program is designed to be more sensitive in identifying milRNAs, meaning it can detect sequences that have less similarity to canonical miRNAs or that have less conserved secondary structures. This increased sensitivity means that it predicts a larger number of potential milRNAs, including sequences that miRDeep2 might not consider as strong candidates.

On the other hand, miRDeep2 is optimized for the prediction of miRNAs in animals and plants, where these sequences follow certain patterns of processing and structure. Thus, this program is known to be more conservative in its predictions, often prioritizing accuracy and confidence in the prediction of milRNAs. This may imply using more stringent criteria, such as more robust secondary structure, better phylogenetic conservation, or higher and more consistent expression in the sequencing data.

6.4 milRNAs Secondary Structures Results

The secondary structure predictions made with RNAstructure and RNAfold generated, for each predicted milRNA, a graphical structural representation file (in .ps –PostScript-- format) alongside with a descriptive file. The format of the descriptive file differs between both programs. RNAstructure works with .ct files, that is, tabulated text files with several columns that describe the connection of each nucleotide to others in the sequence (they provide a linear description of base pairs). On the contrary, RNAfold works with dp.ps files that represent the probability of base pair formation in the form of a dot plot, with points on the grid indicating which nucleotides form pairs and with what probability.

In order to see the differences regarding the graphical representation format, a comparison can be made between *Figure 13* and *Figure 14*, which represent as an example the predicted secondary structure of a same milRNA generated by RNAstructure and RNAfold, respectively. If we look at it in detail, we see that the represented sequence is the same in both programs (as it should be), only differing in the orientation and the use of thymine (RNAstructure) or uracil (RNAfold). Therefore, this is an indicator that both programs are consistent.

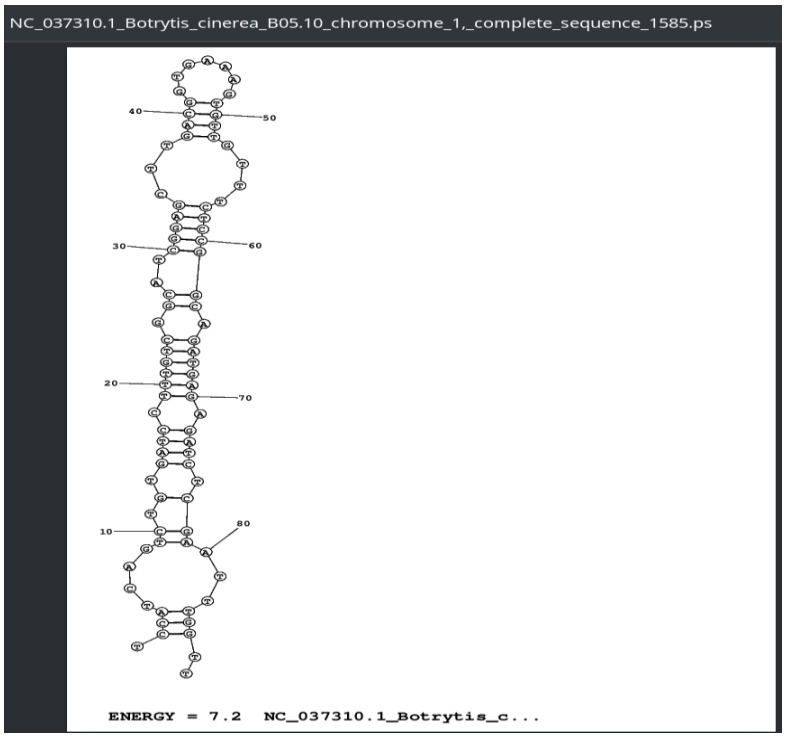


Figure 13. Graphical representation of the predicted secondary structure of a *Botrytis cinerea* milRNA obtained with TurboFold functionality in RNAstructure.

Figure 14. Graphical representation of the predicted secondary structure of a *Botrytis cinerea* milRNA obtained with RNAfold program.

Due to the fact that the prediction programs have generated hundreds of files, the structural information for all the predicted milRNAs can be found in the supplementary material that is presented together with this document (`05_predicted_structures` folders).

After analysing all the graphical representations, we can conclude that those secondary structures coming from the milRNAs predicted by miRDeep2 have a much more robust and consistent structure, similar to those of animal and plant miRNAs. However, structures coming from the milRNAs predicted by milRNApredictor present in many cases weaker connections and some do not resemble canonical miRNAs or are less like. That is why we made a personal selection in a very lax manner, eliminating all the structures that didn't present connections between bases (they had a complete circular aspect) and, thus, couldn't be milRNAs at all.

In *Figure 15* we can see four secondary structures predictions selected in a random manner generated by RNAstructure and coming from milRNAs predicted by milRNApredictor. We can appreciate here the different levels of robustness and miRNA-like similarity commented before.

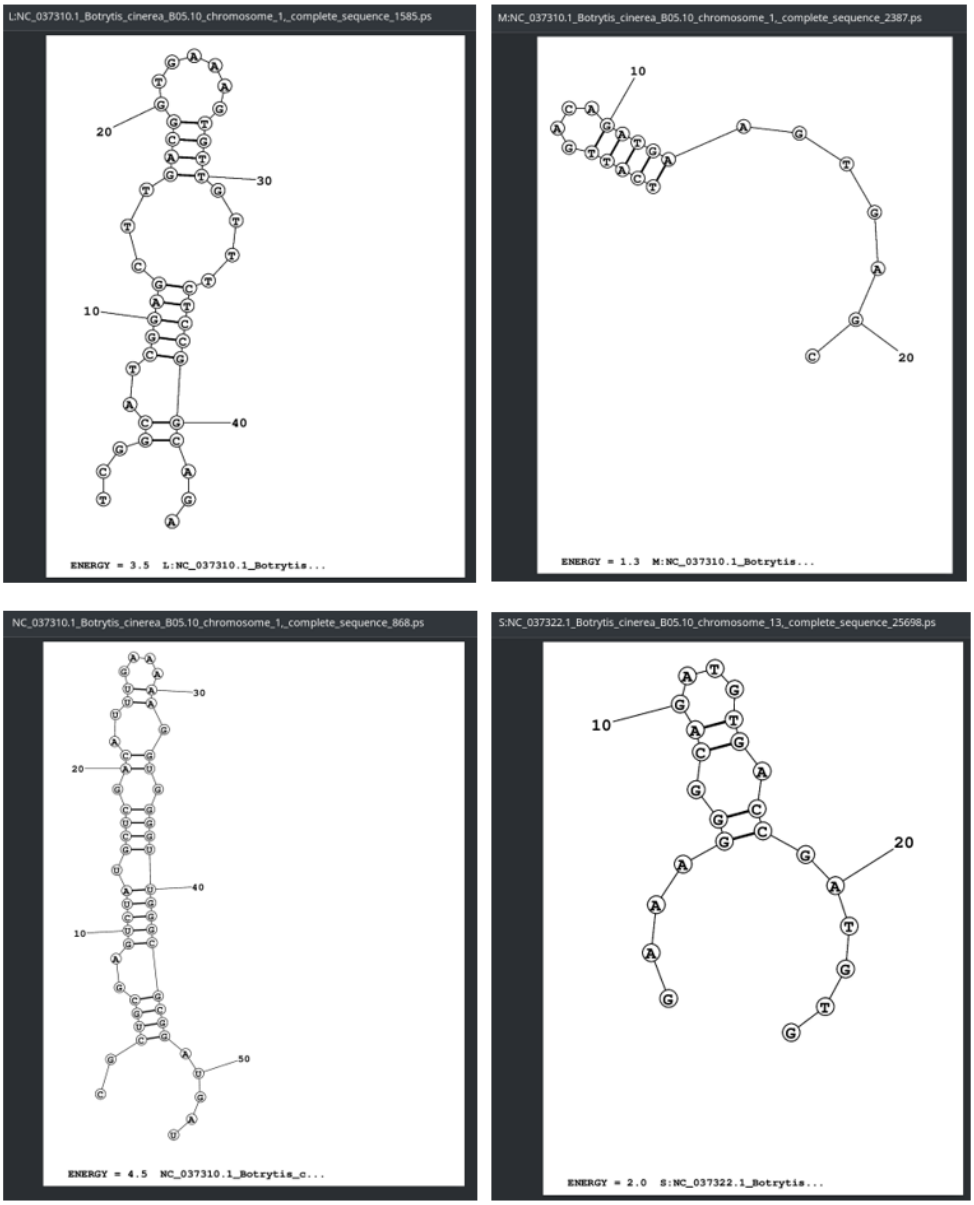


Figure 15. Random selection of four secondary structures generated by RNAstructure. These graphics represent *Botrytis cinerea* milRNAs predicted by milRNApredictor.

There are no remarkable differences between species, beyond the number of predicted structures. That is why we have discussed the results in a generic way, and more oriented towards the behaviour of the prediction programs.

6.5 Target Prediction and Annotation Results

In this section we will explain and discuss the integrated results for all the annotations carried out after the target predictions for each studied species, remarking the most relevant data.

We have created a file called “GO_annotations_integrated.csv” for each pathogen, where all the relevant information about its milRNAs targets has been integrated and presented. The archive contains the following fields:

- Gene → The gene that encodes the mRNA predicted to be target.
- Transcript/s → The target/s itself (mRNAs that have bound with some milRNA).
- Species → Host species where the targets have been sought.
- Description → Functional or annotative description of the gene/transcript.
- Type → Classification of the gene or transcript based on its function or characteristics.
- GO → This field contains Gene Ontology annotations that describe the functions, biological processes, and cellular components associated with a gene. The format established here is [D]GO:NNNNNNN: xxxxxxxxxxx, meaning:
    - [D] → GO domain or category (F for Molecular Function, C for Cellular Component, and P for Biological Process).
    - GO:NNNNNNN → GO term accession (specific GO code for a particular function or process).
    - : xxxxxxxxxxx → GO term name (brief description of the function or process associated with the GO code).
- Enzyme code → Enzyme code/s from the Nomenclature of Enzymes (EC, Enzyme Commission numbers) associated to the gen or transcript. This code classifies enzymes according to the reactions they catalyse.

In order to be able to get more information related to the studied genes, associated metabolic data was added from the annotated enzyme codes with Worksheet application.

This new generated data was grouped in four .csv files called “pathways_data.csv”, one for each organism, with the following annotated fields:

- Pathway → Name of the metabolic or biochemical pathway in which the genes or gene products are involved.
- Seqs in Pathway → Total number of sequences (genes in our case) associated with this specific pathway.
- Enzyme → Specific enzyme activity associated with the sequences in the pathway.
- EnzymeId → Enzyme Commission (EC) number, which uniquely identifies the enzyme's catalytic activity.
- rSeqs → Number of sequences in the dataset that are associated with the specific enzyme activity.
- Seqs → List of the sequences (gene identifiers) that are associated with the enzyme activity within the pathway.
- PathwayId → Unique pathway identifier, derived from KEGG database.
- PathwayImage → Path or link to the image that visually represents the pathway. The image shows how the different genes and enzyme activities are interconnected within the pathway.

All the aforementioned files can be found in the supplementary material, to get a proper general vision of the presented data. Target prediction results are in `06_target_prediction` folders, while both types of .csv annotation documents commented before are in `07_integrated_annotations` folders (pathway images are located inside `Metabolic_pathways` folders).

After gathering all this information, we considered of special importance to seek those metabolic pathways in which there were a greater number of annotated genes involved. This is because this info could be a good indicator of the kind of functions that milRNAs regulate mainly in the selected host species when infected by the desired pathogens. Below are the 10 most active metabolic pathways encountered for each organism, based on the collected data.

Note: “seqs” refers to the total number of different target sequences (genes in our case) involved in the specified metabolic pathway.

- *Botrytis cinerea* (total identified gene targets: 26,282)
    1. mTOR signaling pathway (962 seqs)
    2. PI3K-Akt signaling pathway (960 seqs)
    3. Starch and sucrose metabolism (584 seqs)
    4. Purine metabolism (245 seqs)
    5. Pentose and glucuronate interconversions (236 seqs)
    6. Pyrimidine metabolism (186 seqs)
    7. T cell receptor signaling pathway (184 seqs)
    8. Amino sugar and nucleotide sugar metabolism (173 seqs)
    9. Drug metabolism - cytochrome P450 (153 seqs)
    10. Pyruvate metabolism (152 seqs)

- *Fusarium oxysporum* (total identified gene targets: 20,672)
    1. Starch and sucrose metabolism (377 seqs)
    2. mTOR signaling pathway (357 seqs)
    3. PI3K-Akt signaling pathway (348 seqs)
    4. Pentose and glucuronate interconversions (212 seqs)
    5. Purine metabolism (193 seqs)
    6. Pyrimidine metabolism (141 seqs)
    7. Amino sugar and nucleotide sugar metabolism (135 seqs)
    8. Glycolysis / Gluconeogenesis (114 seqs)
    9. T cell receptor signaling pathway (112 seqs)
    10. Pyruvate metabolism (98 seqs)

- *Phytophthora infestans* (total identified gene targets: 24,499)
    1. mTOR signaling pathway (767 seqs)
    2. PI3K-Akt signaling pathway (759 seqs)
    3. Starch and sucrose metabolism (424 seqs)
    4. Pentose and glucuronate interconversions (216 seqs)
    5. Purine metabolism (199 seqs)
    6. T cell receptor signaling pathway (153 seqs)
    7. Pyrimidine metabolism (140 seqs)
    8. Amino sugar and nucleotide sugar metabolism (134 seqs)
    9. Pyruvate metabolism (123 seqs)
    10. Glycerolipid metabolism (122 seqs)

- *Phytophthora parasitica* (total identified gene targets: 20,131)
    1. mTOR signaling pathway (674 seqs)
    2. PI3K-Akt signaling pathway (669 seqs)
    3. Starch and sucrose metabolism (343 seqs)
    4. Purine metabolism (159 seqs)
    5. Pentose and glucuronate interconversions (136 seqs)
    6. T cell receptor signaling pathway (115 seqs)
    7. Pyrimidine metabolism (114 seqs)
    8. Amino sugar and nucleotide sugar metabolism (111 seqs)
    9. Glutathione metabolism (111 seqs)
    10. Drug metabolism - cytochrome P450 (109 seqs)

The most affected metabolic pathways, that is, the mTOR signalling pathway and the PI3K-Akt signalling pathway, suggest that pathogens may have the ability to interfere with key cellular processes of the host during an infective process, such as growth regulation, cell survival and response to environmental signals [35], [36]. These processes are crucial for the defence of the host and for maintaining the cellular balance, so their alteration could facilitate the infection and spread of the pathogen.

Furthermore, the involvement of the T cell receptor signalling pathway in plants is remarkable because plants do not have T cells like animals. This could reflect that the targets of the predicted milRNAs are involved in similar signalling pathways that regulate the response in plants against stress conditions, environmental factors or as a defence mechanism [37], suggesting manipulation of host defences by pathogens during an attack.

Moreover, pathways involved in starch and sucrose metabolism and pyruvate metabolism are critical for energy generation and nutrient supply. Its alteration could indicate that in an infection the pathogens could be redirecting the host's resources to promote their own proliferation [38], weakening the host in the process.

The fact that the same pathways are affected in the host species by the different pathogens suggests that these pathways are common vulnerabilities in plants. Thus, fungal and fungal-like pathogens may be using similar strategies to infect the different species.

Metabolic pathways with the most genes involved may be key to pathogenesis because they represent crucial aspects of host physiology that pathogens attempt to manipulate. For example, interfering with the mTOR signaling pathway could affect the growth and survival of host cells, facilitating infection.

In summary, these affected metabolic pathways indicate that pathogens could be directing their milRNAs to interfere with key cellular and defense processes in their hosts, weakening them and allowing for more effective infection.

7 Conclusions

This work has had as its main objective the prediction and characterization of microRNA-like RNAs (milRNAs) in four species of plant pathogens of agricultural importance: the fungi *Botrytis cinerea* and *Fusarium oxysporum*, and the oomycetes (fungal-like organisms) *Phytophthora infestans* and *Phytophthora parasitica*. MilRNAs are small fragments of RNA that, despite not being conventional miRNAs, have the ability to regulate gene expression in a similar way. The identification and study of these milRNAs in plant pathogens is a relatively new field of research, but with great potential for understanding how these organisms infect their host plants.

The prediction and characterization of milRNAs in these pathogens is crucial because these small RNAs can play a fundamental role in the regulation of infectious processes. Specifically, milRNAs can interfere with plant defense mechanisms, promoting the infection and spread of pathogens. Understanding how milRNAs work and which targets they regulate in plants can provide new strategies for the control of several diseases, helping to improve crop resistance and reducing associated economic losses.

The process of predicting milRNAs in fungi and oomycetes is complicated for several reasons. Unlike well-characterized miRNAs in other organisms such as plants or animals, milRNAs in fungi and oomycetes have less well-known conservation patterns and can vary significantly between species. In addition, the lack of previous experimental data on milRNAs in these pathogens adds difficulty to the validation of the results. Another challenge is the prediction of the targets of these milRNAs in host plants, since these interactions can be complex and depend on multiple factors.

The results of this study have shown that the predicted milRNAs in *Botrytis cinerea*, *Fusarium oxysporum*, *Phytophthora infestans* and *Phytophthora parasitica* could be involved in the regulation of several critical metabolic pathways in host plants. Pathways such as the mTOR signaling pathway and the PI3K-Akt signaling pathway emerged as major potential targets of these milRNAs, indicating that pathogens could use these milRNAs to alter key processes in plants, such as growth, immune response, and metabolism.

These findings suggest that milRNAs could play a significant role in the pathogenesis of these organisms, and could be exploited as targets for the development of new control strategies. For example, the inhibition of pathogenic milRNAs or the protection of plant targets could represent a way to increase the resistance of crops to infections.

7.1 Impact and Future of Research

This pioneering study in the prediction and characterization of milRNAs in fungi and oomycetes opens the door to new research in the biology of pathogenesis and plant defence. In the long term, the knowledge acquired about milRNAs could be applied in agricultural biotechnology, to develop more resistant crops or to design more effective fungicides and treatments that act on the pathways regulated by milRNAs. It is also hoped that this work will serve as a basis for future studies of milRNAs in other pathogens, thus expanding knowledge about the role of these RNAs in infection and defence in other plant species.

In conclusion, this work has made significant advances in the understanding of milRNAs in fungal and fungal-like plant pathogens, highlighting their importance in the regulation of biological processes essential for infection. Despite the difficulties associated with the prediction and characterization of these small RNAs in little-studied species, the results obtained are promising and could have a considerable impact on crop protection and global food security.

References

[1] “Biotechvana.” Accessed: Jul. 22, 2024. [Online]. Available: https://www.biotechvana.com/

[2] S. T. Hui, H. Gifford, and J. Rhodes, “Emerging Antifungal Resistance in Fungal Pathogens,” *Curr Clin Microbiol Rep*, vol. 11, no. 2, pp. 43–50, Jun. 2024, doi: 10.1007/S40588-024-00219-8.

[3] I. Baruah, G. M. Baldodiya, J. Sahu, and G. Baruah, “Dissecting the Role of Promoters of Pathogen-sensitive Genes in Plant Defense,” *Curr Genomics*, vol. 21, no. 7, pp. 491–503, Jul. 2020, doi: 10.2174/1389202921999200727213500.

[4] X. Yang, L. Zhang, Y. Yang, M. Schmid, and Y. Wang, “miRNA Mediated Regulation and Interaction between Plants and Pathogens,” *Int J Mol Sci*, vol. 22, no. 6, pp. 1–13, Mar. 2021, doi: 10.3390/IJMS22062913.

[5] M. Čáp and Z. Palková, “Non-Coding RNAs: Regulators of Stress, Ageing, and Developmental Decisions in Yeast?,” *Cells*, vol. 13, no. 7, Apr. 2024, doi: 10.3390/CELLS13070599.

[6] L. Li, S. S. Chang, and Y. Liu, “RNA interference pathways in filamentous fungi,” *Cell Mol Life Sci*, vol. 67, no. 22, p. 3849, Nov. 2010, doi: 10.1007/S00018-010-0471-Y.

[7] P. Gowda, P. H. Reddy, and S. Kumar, “Deregulated mitochondrial microRNAs in Alzheimer’s disease: Focus on synapse and mitochondria,” *Ageing Res Rev*, vol. 73, Jan. 2022, doi: 10.1016/J.ARR.2021.101529.

[8] Z. A. Syeda, S. S. S. Langden, C. Munkhzul, M. Lee, and S. J. Song, “Regulatory Mechanism of MicroRNA Expression in Cancer,” *Int J Mol Sci*, vol. 21, no. 5, Mar. 2020, doi: 10.3390/IJMS21051723.

[9] M. A. Mori, R. G. Ludwig, R. Garcia-Martin, B. B. Brandão, and C. R. Kahn, “Extracellular miRNAs: From Biomarkers to Mediators of Physiology and Disease,” *Cell Metab*, vol. 30, no. 4, pp. 656–673, Oct. 2019, doi: 10.1016/J.CMET.2019.07.011.

[10] K. Bi, Y. Liang, T. Mengiste, and A. Sharon, “Killing softly: a roadmap of Botrytis cinerea pathogenicity,” *Trends Plant Sci*, vol. 28, no. 2, pp. 211–222, Feb. 2023, doi: 10.1016/J.TPLANTS.2022.08.024.

[11] M. Otto *et al.*, “Botrytis cinerea expression profile and metabolism differs between noble and grey rot of grapes,” *Food Microbiol*, vol. 106, Sep. 2022, doi: 10.1016/J.FM.2022.104037.

[12] C. Srinivas *et al.*, “Fusarium oxysporum f. sp. lycopersici causal agent of vascular wilt disease of tomato: Biology to diversity– A review,” *Saudi J Biol Sci*, vol. 26, no. 7, pp. 1315–1324, Nov. 2019, doi: 10.1016/J.SJBS.2019.06.002.

[13] S. C. Whisson, P. C. Boevink, S. Wang, and P. R. Birch, “The cell biology of late blight disease,” *Curr Opin Microbiol*, vol. 34, pp. 127–135, Dec. 2016, doi: 10.1016/J.MIB.2016.09.002.

[14] K. A. Midgley, N. van den Berg, and V. Swart, “Unraveling Plant Cell Death during Phytophthora Infection,” *Microorganisms*, vol. 10, no. 6, Jun. 2022, doi: 10.3390/MICROORGANISMS10061139.

[15] I. Matsiakh and A. Menkis, “An Overview of Phytophthora Species on Woody Plants in Sweden and Other Nordic Countries,” *Microorganisms*, vol. 11, no. 5, May 2023, doi: 10.3390/MICROORGANISMS11051309.

[16] A. Kaur, A. Arora, and M. S. Hunjan, “ Association of Phytophthora nicotianae var. parasitica with citrus foot rot in Punjab ,” *Agricultural Research Journal*, vol. 53, no. 4, p. 609, 2016, doi: 10.5958/2395-146X.2016.00125.3.

[17] Y. Iftikhar *et al.*, “Citrus Gummosis: A Global Threat to Citrus Production and Quality-Distribution, Diagnosis, and Management Strategies,” *Plant Protection*, vol. 7, no. 2, pp. 351–359, Aug. 2023, doi: 10.33804/PP.007.02.4727.

[18] A. I. Hafez *et al.*, “Client Applications and Server-Side Docker for Management of RNASeq and/or VariantSeq Workflows and Pipelines of the GPRO Suite,” *Genes (Basel)*, vol. 14, no. 2, Feb. 2023, doi: 10.3390/GENES14020267.

[19] H. Shi and X. Xu, “Learning the Sequences Quality Control of Bioinformatics Analysis Method,” 2016, doi: 10.2991/ICEEMT-16.2016.90.

[20] “Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data.” Accessed: Aug. 27, 2024. [Online]. Available: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[21] M. Martin, “Cutadapt removes adapter sequences from high-throughput sequencing reads,” *EMBnet J*, vol. 17, no. 1, p. 10, May 2011, doi: 10.14806/EJ.17.1.200.

[22] R. Schmieder and R. Edwards, “Quality control and preprocessing of metagenomic datasets,” *Bioinformatics*, vol. 27, no. 6, pp. 863–864, Mar. 2011, doi: 10.1093/BIOINFORMATICS/BTR026.

[23] “Bowtie: An ultrafast, memory-efficient short read aligner.” Accessed: Aug. 27, 2024. [Online]. Available: https://bowtie-bio.sourceforge.net/index.shtml

[24] M. R. Friedländer, S. D. MacKowiak, N. Li, W. Chen, and N. Rajewsky, “miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades,” *Nucleic Acids Res*, vol. 40, no. 1, pp. 37–52, Jan. 2012, doi: 10.1093/NAR/GKR688.

[25] Y. Yao, H. Zhang, and H. Deng, “milRNApredictor: Genome-free prediction of fungi milRNAs by incorporating k-mer scheme and distance-dependent pair potential,” *Genomics*, vol. 112, no. 3, pp. 2233–2240, May 2020, doi: 10.1016/J.YGENO.2019.12.019.

[26] “Worksheet.” Accessed: Aug. 27, 2024. [Online]. Available: https://gpro.biotechvana.com/tool/Worksheet/manual/overview

[27] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, “CD-HIT: accelerated for clustering the next-generation sequencing data,” *Bioinformatics*, vol. 28, no. 23, pp. 3150–3152, Dec. 2012, doi: 10.1093/BIOINFORMATICS/BTS565.

[28] S. E. Ali, A. Mittal, and D. H. Mathews, “RNA Secondary Structure Analysis Using RNAstructure,” *Curr Protoc*, vol. 3, no. 7, p. e846, Jul. 2023, doi: 10.1002/CPZ1.846.

[29] Z. Tan, Y. Fu, G. Sharma, and D. H. Mathews, “TurboFold II: RNA structural alignment and secondary structure prediction informed by multiple homologs,” *Nucleic Acids Res*, vol. 45, no. 20, pp. 11570–11581, Nov. 2017, doi: 10.1093/NAR/GKX815.

[30] I. L. Hofacker, “Vienna RNA secondary structure server,” *Nucleic Acids Res*, vol. 31, no. 13, p. 3429, Jul. 2003, doi: 10.1093/NAR/GKG599.

[31] R. Lorenz *et al.*, “ViennaRNA Package 2.0,” *Algorithms for Molecular Biology*, vol. 6, no. 1, pp. 1–14, Nov. 2011, doi: 10.1186/1748-7188-6-26/TABLES/2.

[32] A. Rueda *et al.*, “sRNAtoolbox: an integrated collection of small RNA research tools,” *Nucleic Acids Res*, vol. 43, no. Web Server issue, p. W467, Jul. 2015, doi: 10.1093/NAR/GKV555.

[33] Á. L. Riffo-Campos, I. Riquelme, and P. Brebi-Mieville, “Tools for Sequence-Based miRNA Target Prediction: What to Choose?,” *Int J Mol Sci*, vol. 17, no. 12, Dec. 2016, doi: 10.3390/IJMS17121987.

[34] D. Smedley *et al.*, “BioMart – biological queries made easy,” *BMC Genomics*, vol. 10, p. 22, Jan. 2009, doi: 10.1186/1471-2164-10-22.

[35] S. Rashidi *et al.*, “miRNAs in the regulation of mTOR signaling and host immune responses: The case of Leishmania infections,” *Acta Trop*, vol. 231, Jul. 2022, doi: 10.1016/J.ACTATROPICA.2022.106431.

[36] S. Rashidi *et al.*, “The host mTOR pathway and parasitic diseases pathogenesis,” *Parasitol Res*, vol. 120, no. 4, pp. 1151–1166, Apr. 2021, doi: 10.1007/S00436-021-07070-6.

[37] J. Kumar, A. Ramlal, K. Kumar, A. Rani, and V. Mishra, “Signaling Pathways and Downstream Effectors of Host Innate Immunity in Plants,” *Int J Mol Sci*, vol. 22, no. 16, Aug. 2021, doi: 10.3390/IJMS22169022.

[38] P. Kanwar and G. Jha, “Alterations in plant sugar metabolism: signatory of pathogen attack,” *Planta 2018 249:2*, vol. 249, no. 2, pp. 305–318, Sep. 2018, doi: 10.1007/S00425-018-3018-3.