Universidad Zaragoza
1542

Instituto Universitario de Investigación
Biocomputación y Física
de Sistemas Complejos
Universidad Zaragoza

Master's dissertation:

Multi-omics characterization of the function of virR protein in cell wall remodeling in *Mycobacterium tuberculosis*

Author:
Jorge Bertol Faure
Directors:
Joaquín Sanz Remón
Ignacio Marchante Hueso

Academic year 2022-2023
MASTER'S DEGREE IN BIOPHYSICS AND QUANTITATIVE BIOTECHNOLOGY
Universidad de Zaragoza

Index

List of figures

List of tables

Glossary

EVs: Extracellular Vesicles

M.tb.: *Mycobacterium tuberculosis*

WT: Wild Type

RNA-seq: RNA Sequencing

NGS: Next Generation Sequencing

MS: Mass Spectrometry

LFQP: Label Free Quantitative Proteomics

PCA: Principal Component Analysis

PC: Principal component

M[N]AR: Missing [Not] at Random

DEA: Differential Expression Analysis

DE: Differentially Expressed

logFC: logarithm of Fold Change (base 2)

FWER: Family-Wise Error Rate

FDR: False Discovery Rate

GO: Gene Ontology

KEGG: Kyoto Encyclopedia of Genes and Genomes

PDIM: phthiocerol dimycocerosate

WC: Whole Cell

KO strain: Knockout strain

C strain: Complemented strain

MLE: Maximum Likelihood Estimation

Abstract

VirR, a LytR_C domain containing protein, has been linked to the regulation of production of extracellular vesicles (EVs) in *Mycobacterium tuberculosis* by maintaining cell wall integrity. To test its regulatory role in vesiculogenesis, proteomic and transcriptomic profiles from a mutant strain of *M. tuberculosis* with an inactive phenotype of the VirR protein (VirR-KO) were obtained, as well as from the wild-type strain H37RV. The VirR KO mutant, as ascertained through previous results of our collaborators, presented an abnormal cell wall morphology linked to higher permeability, lower virulence and an increased production of extracellular vesicles.

Transcriptomic data was collected from whole cell lysates from each strain, revealing major differences concerning host defense and host-induced stress responses, as well as gene regulation, and metal ion import and export, which is coherent with the divergent virulence and secretory profiles of H37RV and the VirR-KO mutant. Furthermore, these analyses also revealed divergent expression levels for genes involved in mRNA translation, stabilization, and metabolism; thus pointing to the existence of relevant post-transcriptional regulatory mechanisms mediated by VirR.

To test that hypothesis, proteomic analyses were performed on whole cell lysates and isolated extracellular vesicles from each strain, revealing largely disjoint sets of differentially expressed proteins with respect to mRNAs, further pointing to the relevance of post-transcriptional mechanisms in the regulatory role of VirR. Differential expression analysis of the proteins revealed greater differences between extracellular vesicles from different strains than between whole cell lysates, indicating a primary role for VirR in the proteomic enrichment profile of EVs.

Taken together, our results highlight a relevant regulatory role for virR that leans both on transcriptional and post-transcriptional mechanisms, which significantly mediates the amount and function of EVs secreted by *Mycobacterium tuberculosis*.

1. Introduction and main goals

Tuberculosis, the disease caused by *Mycobacterium tuberculosis* (M.tb.), is the main cause of death from a single pathogen worldwide. It is estimated that one third of the world's population is infected with the disease, either with an active infection or with a latent one. In 2020, 1.5 million people died because of the disease (1).

The main cell type targeted by M.tb. cells are macrophages, one of the first barriers of defense of the human immune response. This infection involves the modulation of the normal progression of the macrophage activity by the pathogen and avoiding the development of a localized immune response which could activate the host cell (2).

Despite the existence of a vaccine for the infection and the availability of an antibiotic treatment, they have several disadvantages. First of all, drug-resistant tuberculosis cases are increasing each year, which means that it is necessary to further research on new strategies to treat the disease (1). Moreover, the current Bacille Calmétte-Guerin vaccine (BCG), while having been very helpful in the battle against tuberculosis, especially in the most morbid forms of extra-pulmonary, pediatric TB, does not have a comparable effectivity against the pulmonary forms of the disease, which are responsible of the bulk of new infections, mainly in adult age groups (3). These facts highlight the importance of increasing research on the topic of tuberculosis, in order to discover novel approaches towards diagnosis, prevention and treatment of this disease.

Recent findings (4) highlight the importance of EVs as key elements for cell-to-cell communication, and modulation of host responses by M.tb., even though these have been widely disregarded in the literature when it comes to bacteria. EVs have been proposed as a potential target to develop new diagnostic tools, drugs and vaccines, and a deeper insight about their role is instrumental to further characterize the M.tb. infection cycle.

This Master Thesis project is devoted to shed light on the function of these vesicles, the way they are created and their implications on the biological processes of M.tb.. In order to achieve that, a combination of transcriptomics and proteomics data has been analyzed to assess the differences between a wild type strain of M.tb. and a mutant where a gene known to be key in the regulation of generation of vesicles (virR) (4) is knocked out.

2. Biological background and concepts

Before delving into the analyses performed, I deem of importance to stablish a series of concepts so that further conclusions are easier to follow.

2.1. M.tb. cell cycle

M.tb. infection has several peculiarities that allow it to be one of the most successful pathogens. It starts with the internalization of M.tb. cells by alveolar macrophages that the bacteria have reached after being inhaled by the infected host.

M.tb. invades the phagosomal compartment of macrophages, which is the main weapon of these immune cells against pathogens. This hostile environment, characterized by an acidic composition and oxygen and nitrogen reactive species, is nonetheless often survived by the bacteria, and its ability to do so is actually crucial for the development of the infection.

When a macrophage is able to complete a successful phagocytosis event, phagosomes are fused with other cell compartments and finally deliver the internalized pathogens to a structure called lysosome, which has the function of completely eliminating pathogens. But in the case of mycobacteria, and even though the conditions are extremely hostile for the survival of living organisms, the pathogen is often able to resist this environment and even manipulate the host cells to generate a non-hostile location for its own growth. M.tb. cells have been observed to prevent apoptosis of macrophages (which is often bactericidal) and in turn induce necrosis, which is a non-regulated way of cell death that allows M.tb. cells to safely invade other cells (5).

The *trick* behind the ability of M.tb. cells to resist the dangers of the environments that they must face is the capacity to assume a dormant state, becoming apparently quiescent and being able to survive without activating host immune response for years, while maintaining a metabolically that is partly inactive. It is known that M.tb. cells in this state maintain replication, but in a much lower rate than active bacteria, mainly using fatty acids as carbon source (6). From a tissular perspective, macrophages harboring dormant bacteria often trigger the formation of granuloma around them (Figure 1). The granuloma is a microcellular structure conformed by macrophages and other immune system components generated by the immune system when an infection cannot be fully cleared, but has to be contained within a localized, enclosed structure.

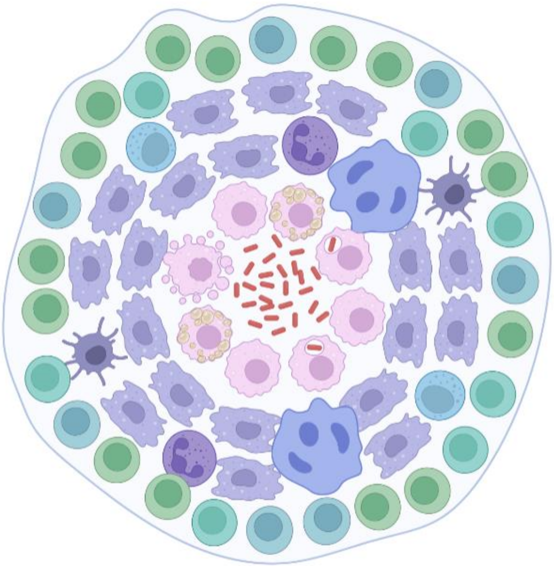


Figure 1. Schematic representation of a granuloma.

M.tb. exploits this situation and uses the granuloma to survive during long periods of time, out of grasp of potentially bactericidal additional host's immune responses, and even antibiotics. Eventually, dormant M.tb. bacteria can become active again and reinitiate the infection, which is usually triggered by a decrease in host immune activity. M.tb. cells in this state are difficult to detect, which raises a need for the correct diagnosis of a dormant infection to increase the capacity to treat it before pathogen activation (7).

2.2. Mycobacterial cell wall

The mycobacteria family is characterized by a modified version of the gram-positive cell wall. The typical structure of a gram-positive cell wall consists of a lipid cell membrane followed by a thick layer composed of what is known as peptidoglycan, which is a compound formed by the union of peptides and sugars (Figure 2). This structure confers rigidity and resistance to the bacteria, and also helps in maintaining their shape and integrity (8).

The mycobacterial cell wall global structure is similar to that of gram-positive bacteria but has several specific differences. The peptidoglycan of mycobacteria features a layer of a polysaccharide called arabinogalactan, which is also connected to the outer part of the wall, which is conformed by a series of mycobacteria-specific lipids called mycolic acids. These last two characteristic mycobacterial structures have been observed to have a key role in the pathogenesis of mycobacteria and also their survival. Historically, this involvement has motivated the study of a number of cell-wall related proteins as important targets for drug development against mycobacteria (9).

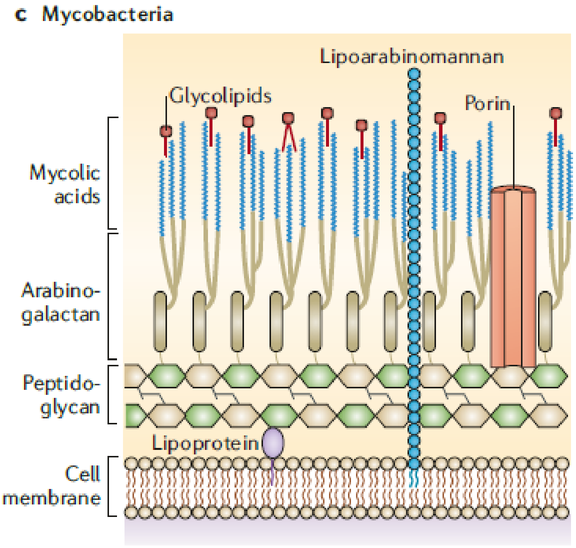


Figure 2. Diagram of the structure of the mycobacterial cell wall (10)

2.3. Host-pathogen cross-talk in M.tb.. infection

During an infectious process, one of the determinant factors of the outcome of the infection are the interactions that host, and pathogen establish. Pathogens are often adapted to respond to host defenses so as to elude them and to counteract these efforts by generating virulence factors, which can have a plethora of different functions ranging from metabolic disruption to toxicity. As one may expect, host cells also respond to the pathogen reactions, resulting in a metabolic and signaling *argument of sorts* between pathogen and host, which result will determine the fate of the infection. This phenomenon receives the name of “cross-talk”(11).

Cross-talk is especially important in the case of intracellular pathogens. In case of M.tb., there is a direct competition for resources inside the host cell, and due to the characteristics of the infection, intracellular pathogens have evolved mechanisms to reduce to the maximum the energy consumption while optimizing responses to host attacks. This is achieved, for example, by utilizing by-products of essential pathways (such as fermentation) to generate a desired response on the host after releasing them to its cytoplasm. Furthermore, speed in cross-talk responses is key, as changes in metabolism of the antagonistic cell are very rapidly detected either by the host or the pathogen, yielding an especially tight co-regulatory phenomenon (11).

2.4. Extracellular Vesicles

Recently, the scientific community has focused on researching extracellular vesicles (EVs), which have been described as an important factor in the cross-talk between all kinds of cells (12).
EVs consist of lipid bilayers that form spheres with a size range from 20 nm to 500 nm in diameter. They are found in eukaryotic and prokaryotic cells alike, which suggests that vesicle-mediated transport is a widely conserved process across life taxa (10).

Study of gram-positive and mycobacterial EVs has suffered from a lack of interest until the last few years, due to the assumption that they could not exist, for gram-positive bacteria and mycobacteria have a thick peptidoglycan wall outside of their cell membrane, with no outer membrane such as the one that gram-negative bacteria have. It was thought that this wall did not allow EVs to be released, and thus research about the topic was no deemed important. In 2007, the first mycobacterial EVs were isolated, which started their study for different organisms that were previously thought unable to produce them (10).

While no certain answer to how EVs are produced is known, three compatible hypotheses are supported in the literature about this topic (Figure 3):

- EVs may be forced through the wall by turgor pressure after release from the plasma membrane.
- EVs production may be facilitated by specific cell-wall modifying enzymes.
- EVs may pass through channels, facilitated by their plasticity. They may deform and thus be able to pass through channels narrower than their usual size (10).

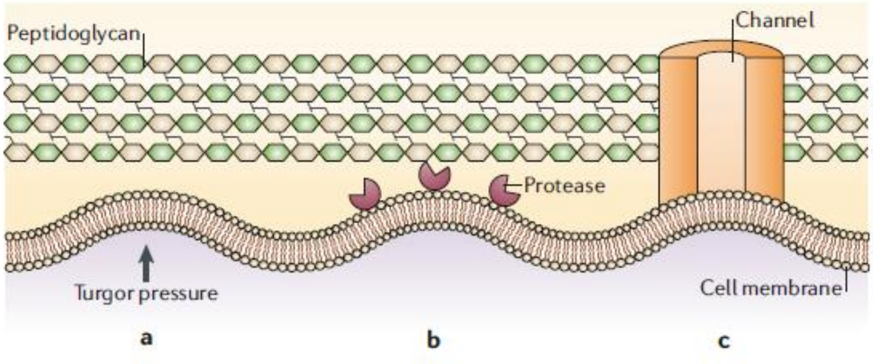


Figure 3. Diagram of the three working hypotheses related to EV formation. a) Turgor pressure. b) Facilitation by cell-wall modifying enzymes. c) Channel-facilitated secretion.(10)

There is a wide variety of molecules that have been found in EVs, including polysaccharides, proteins, and nucleic acids. The released cargo has been associated to several activities, ranging from antibiotic activity to toxin release and virulence factors. An important hypothesized role of

EVs is the transference of DNA containing information for antibiotic resistance, and also interspecies communication between bacteria (10).

2.5. EVs in M.tb. and their interest

We have already highlighted the importance of cross talk between pathogen and host, and *M. tuberculosis* is no exception. Its cellular cycle requires efficient sensing of host activity and also the delivery of virulence factors to facilitate the infectious process. One of the hypothesized mechanisms by which they are able to deliver these molecules is through EVs. The way that bacterial cells select the cargo that will be delivered to the exterior is still unknown but advances on research of the regulation of EVs production in M.tb. have been recently made. EV-producing pathogenic bacteria have been related to a more effective proliferation and virulence than non-producing ones, suggesting a significant role of these structures in M.tb. virulence and survival.

The previously mentioned hypothesis of EV production in bacteria are also applied to M.tb. cells. This vesiculogenesis has been observed to be regulated by several processes and environmental conditions. When M.tb. meets an iron deficienct environment, EV production is increased, marking iron limitation as a regulator of vesiculogenesis. Another important regulator that has been found in mycobacteria is the VirR protein, which is of utmost interest for the analyses made in this Master thesis (13).

Many studies have been made aiming at exploiting extracellular vesicles either as infection markers -for diagnosis, or vaccine development-, or as actionable systems -for the development of therapeutic tools-. Studies in their protein cargo have revealed that M.tb. EVs contain unique proteins found among a majority of active TB patients, which hints at a possible diagnostic utility for these vesicles. Not only that, but the detection of extracellular vesicles in patients with presence of latent M.tb. have also showed that EVs at this stage may have a specific cargo which could help in detection of latent tuberculosis, a very important necessity due to the difficulty for its detection with current methods (14).

Another important utility of EVs is the production of vaccines that use them to generate immunity. Their nature offers several advantages to this end. First of all, EVs are released to the extracellular medium and distribute through all body fluids, which can help to maximize the immune response of a hypothetical EV-based vaccine. Second, their small size allows them to distribute easily even through the blood-brain layer, allowing for immunity generation against pathogens with the ability to cross it (4). Their surface composition is similar to that of the pathogens from which they come. Another important feature is their adjuvant characteristic. Adjuvants are compounds that are used in vaccines to increase the immunogenic response. As EVs can also act as adjuvants, this would make them a great choice for more easily produced vaccines. Last but not least, EV vaccines are a safer alternative to inactivated vaccines such as the BCG vaccine for M.tb., as what is introduced inside the patient is not an inactivated cell but rather a component with no infectious capacity. M.tb. EVs have been proved to generate immunity in mice (15), so the path towards an EV vaccine is open, but research is still necessary. Because of this, a better comprehension of EV nature and regulation mechanisms is important.

2.6. VirR

VirR is a 16’9 kDa protein from M.tb. involved in genetic regulation that contains a LytR_C domain (structure, as predicted by Alphafold, in Figure 4), which is related to enzymes that participate in cell wall processes (16) . Recent studies have identified virR as a regulator of EV formation and immune modulation (17). This protein contains a highly hydrophobic region, which has been associated to a possible binding to hydrophobic surfaces such as the cell membrane. Several binding partners have been found, all of which have a membrane-binding predicted activity, which further supports its involvement in membrane processes and vesicle formation (18).

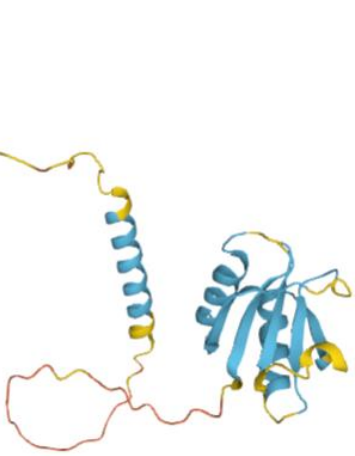


Figure 4. Alphafold prediction of the structure of VirR. Source: Uniprot.

VirR-KO mutant M.tb. bacteria have been associated to a higher sensitivity towards some antibiotics (16), which has been related to a possible malfunction in cell wall biogenesis. This protein has also been related to regulation of immunomodulation of macrophages. VirR-deficient M.tb. cells have shown a greater rate of activation of their host macrophages when comparing them to Wild-Type (WT) cells. An increase of extracellular vesicle formation in virR deficient cells has also been reported, which is thought to be related to a deficiency in structure of the cell wall. To sum up, the proposed role of VirR is the regulation of immunomodulation, release of EVs and cell-wall biogenesis (19). For all these reasons, VirR seems to be an interesting target in M.tb. EVs research. The project of this master Thesis studies in fact a mutant strain where VirR is absent and compares it to a WT strain.

3. Biocomputational background and concepts

Now that the different biological concepts involved in this project have been clarified, we may move on to stablish some terms related to the computational analysis performed and the techniques involved in this project.

3.1. RNAseq

RNA-seq is a commonly used technique in biosciences research that is conducted to study the presence and amount of RNA in cell samples, at a specific point of time determined by the moment of the extraction and purification of the tissue samples. In our case, this technique leans on Next Generation Sequencing (NGS), a technology that allow to sequence the massive amount of different RNAs while also maintaining the possibility of quantifying their relative amount with respect to the whole RNA extract.

3.1.1. Next Generation Sequencing

Next-Generation sequencing is a term that refers to a family of technologies of nucleic acids sequencing which has unlocked the possibility of conducting RNA-seq analysis on whole transcriptomes of cells for a reduced price. This term encompasses many different techniques which have in common their speed and capacity to offer a high throughput. One of the most widely used experimental platforms for NGS is Illumina, which is especially useful for RNA seq experiments where the final goal is the quantification of gene expression in species whose transcriptome has been previously assembled and it is thus well known.

Illumina sequencing requires a first step of library preparation. This step consists of converting the extracted RNAs into cDNAs using a reverse transcription process. As no amplification is made at this point, the amount of cDNA that is produced is proportional to the amount of RNA that was imputed. After that, cDNA is fragmented and attached to small DNA sequences called adaptors, which are required for sequencing, and sample demultiplexing. The mechanism of the Illumina sequencing technique involves the attachment of these cDNA fragments to a surface (by adaptors' hybridization) where they will be amplified via PCR in a way that generate clusters of cloned sequences across a flow cell surface. After the amplification step, sequencing is done *in situ*, and the data of the sequences is collected. The amount of cDNAs of a certain gene is proportional to the input quantity of that DNA, allowing estimating, quantitatively, the amount of RNA mapping to each gene or transcript in each sample.

3.2. Mass Spectrometry-Based Label-Free Quantitative-Proteomics

Mass Spectrometry-Based Label-Free Quantitative-Proteomics is a protein-quantification technique that allows to detect and differentiate many proteins using mass spectrometry (MS) without the need for previous biological labelling of proteins from different samples. This technique has a disadvantage when compared to labeled techniques, as these last ones allow to analyze all samples using the same gel, which eliminates potential gel-to-gel variations, but on the other hand offers the advantage of not depending on gel characteristics and reducing ambiguity in protein quantification (20).

There exist many Label Free Quantitative Proteomics (LFQP) methods, but all have in common the following steps:

- Protein extraction, reduction, alkylation, and digestion
- Sample preparation by liquid chromatography separation and mass spectrometry analysis
- Data analysis, which includes peptide identification, quantification, and statistical analysis.

When we deal with label-free proteomics, these steps are performed in each sample separately. Quantification in this technique is mainly done following two different methods:

- Relative quantification by peak intensity: This method takes advantage of the fact that intensity of detection and peptide abundance are strongly correlated. Abundance of proteins in the same sample can easily be compared by comparing peak intensity, but normalization is required before comparison between samples.

- Relative quantification by spectral count: Through this method, one compares the number of identified MS/MS spectra from the same protein in each of the multiple datasets. This method assumes that an increase in protein abundance will result in an increase in the number of its proteolytic peptides, and vice versa. This correlation has been demonstrated, and allows a quick, reliable and direct comparison between protein abundances in the same sample.

3.3. Principal Component Analysis

Principal Component Analysis (PCA) is often used to check the characteristics of the samples from a dataset and their relation before differential expression analysis. PCA is a statistical technique that is mainly used on datasets with a large number of dimensions, e.g., datasets with a lot of variables for each sample. PCA reduces the dimensions of the data by performing a linear transformation to obtain new non-correlated variables called “principal components”. Each of them corresponds to a projection of the data (defined by a linear combination of the variables), each of which capture the biggest amount of variability in the data, once the previous ones have been regressed out. The components are not correlated between each other (i.e., they are orthogonal) and are ordered decreasingly according to the fraction of variance from the original data each of them explains. As the first PCs are bound to capture a big part of the data heterogeneity, they are useful to represent data with fewer dimensionality, to conduct QC at the level of samples, and to check similarity/dissimilarity between groups of them.

Typically, the projection of the data in the first two components is represented by plotting the PC1 and PC2 projections as points. By doing this, one can qualitatively inspect if there is or not a clear separation between groups of samples. This way, an exploratory analysis of the data can be easily done regardless of the dimensionality of the set. It is important to note that this method, although useful, is not enough to describe all the characteristics of a dataset but can help in understanding if the samples are different in behavior and, if that is the case, how different they are.

3.4. Data treatment in proteomics

3.4.1. Normalization

Before we can compare the protein amount readouts between conditions in a proteomic dataset, we need to pre-treat the data as to fairly assess the differences through a normalization transformation. This step is done prior to differential expression analysis, to allow correct comparisons between samples and proteins, aiming to reduce the bias associated with technical effects.

Many different normalization methods exist for label-free quantitative proteomics. Several of them are directly derived from methods that were developed to work with DNA microarrays, as the assumptions they lean on can be generalized to mass spectrometry data (21).

3.4.2. Imputation

In label-free mass spectrometry experiments, the appearance of missing values is a common issue that needs to be dealt with. Missing values often appear as zeroes in data, but they do not necessarily mean that there is an absence of the specific peptide that is being quantified. Depending on their nature, missing values can be classified as follows:

-Missing at Random (MAR): These missing values are a consequence of stochasticity or technical issues during the experiment. They do not represent absence of a peptide, but a failure at collecting its corresponding data caused by either stochastic fluctuations or the propagation of errors during the experiment.

-Missing Not at Random (MNAR): These missing values represent true absence of signal of a peptide, due to a very low concentration or lack of presence (22).

Missing value imputation is a step whose goal is to assign reliable values to missing data in order to avoid the loss of important information and to mitigate data sparsity. Methods for MAR missing values are abundant and well-developed, as they deal with an issue that is present in many fields.

An ideal procedure for missing value imputation would be to differentiate missing values in data according to their nature, and then use different missing value imputation procedures depending on their type. This is known as mixed imputation. Unfortunately, this is not an easy task, as there is no direct way of assigning categories to missing values without assuming certain risk of bias. Because of this, another option is to use methods that can account for the possible presence of both types of data, borrowing information from the dataset to assign a larger or smaller value to each individual missing value. To this date, there is still no consensus about which imputation method is best, and it is usually considered to be dataset dependent.

3.4.3. Differential expression analysis

The concept of differential expression analysis (DEA) refers to a type of statistical analysis that aims to detect differences in the expression of a certain biological agent, such as genes or proteins, across the values of variables related to samples attributes, either factorial or continuous. As we are interested in inspecting differences between conditions, we simplify and consider only factorial variables in the explanation. A factorial variable is defined as a categorical variable according to which samples can be classified into a number n of possible conditions. Each of the conditions is commonly referred to as a variable level. DEA has been mainly applied to transcriptomic data but can also be applied to data from proteomics experiments.

In order to carry out DEA, one needs to declare an experimental design matrix. This matrix encodes the comparisons that will be tested for size and direction of the difference between gene expression/protein abundance. The rows of the matrix correspond to the samples and the columns to the levels of the factorial variables to be tested. Each sample has a value assigned to each level, being it 1 if the condition associated to the factor level applies to the sample and 0 otherwise. This is done for all levels of each factorial variable modelled except one: the reference level, in such a way that each column introduced in the design matrix will capture the differences between the expression of each included level, and the reference one. Typically, this technique involves the use of a linear model per feature in the data, yielding the following mathematical representation of the procedure, when only a unique factorial variable with n levels is to be modelled:

$$E_i^j = \beta_{i,o} + \sum_{k=1}^{n-1} \beta_{i,k} * c_k^j + \varepsilon_i^j$$

E_i^j represents the expression value for a given feature i, in a certain sample j.
The terms colored in green are the ones that correspond to the part of the expression that is explained by the model, and this is the part that depends on the experimental design:
$\beta_{i,o}$ represents the expected expression level in the reference condition for feature j (Typically referred as intercept).
$\beta_{i,k}$represents the size effect for a gene i between the condition k of the design and the reference one.
c_k^j is a dummy variable representing a column of the design, and its value depends on to which level of the factorial variable the j-th sample belongs to. It takes only two values: 1 if the current level is level k, and 0 otherwise. This variable allows to only apply the $\beta_{i,k}$ coefficient on its corresponding condition, in such a way that $\beta_{i,k}$ captures the difference in expression foreseen by the model between the reference level and the k-th level, for gene i.
ε_i^j represents the part of the expression that is not explained by the model. These values are called the residuals and are the difference between the real value of expression and the estimated value.

The identity and meaning of β coefficients and the dummy variables are determined by the experimental design, as they depend on what variables are present and tested, and, specially, on their reference values. The values of these coefficients are often referred as logFCs, which is short of logarithm of fold changes, as the differences between transformed abundances occur on a logarithmic space where subtraction corresponds to multiplicative changes in the untransformed framework. In this sense, these models are equivalent, and often referred to as generalized linear models with a logarithmic link.

After the experimental design is defined, DEA methods compute the logFCs corresponding to each contrast for all features, and then estimate the standard errors for each one of them. These logFCs, and their standard errors are finally used to conduct hypothesis contrast to evaluate the compatibility of each estimated effect with the null hypothesis of no-effect. From these contrasts, one can compute the associated p-values for each term, and gene, which act as nominal significance variables to assess whether a given feature (gene or protein) is deemed as differentially expressed or not for a particular contrast.

3.4.4. Multiple testing correction

As with all statistical tests, p-values in DEA are used as a reference for selecting statistically significant changes in expressed proteins and genes between conditions. A common issue of -omics data is that, due to the size of the datasets, some features will be falsely considered significant during DEA (and other procedures) by chance, just as a result of multiple testing, which is of course not a desirable outcome. Because of this, multiple testing correction is necessary to produce estimates of the expected rates of false positives that are found, to ensure that reported statistical significance thresholds are associated to known, and low enough rates of false positives (23).

Multiple test correction strategies are often classified in those that aim to correct for the Family-Wise Error Rate (FWER) or False Discovery Rate (FDR):

-FWER refers to the probability of making at least one false discovery. Methods that focus on reducing this rate are stringent, trading statistical power for a lower number of false discoveries.

-FDR refers to the expected proportion of false discoveries, taking as reference all discoveries under a given nominal significance level. Methods that control FDR are less stringent, as they focus on estimating the total number of false positives within a set of discoveries, and not the probability that such set contains at least 1.

In -omics analyses, where the outcome of models are sets of genomic features (genes, or proteins), that typically capture, as a whole, the nature of the biological processes and pathways associated to a given condition, the most common approach consists of calculating FDRs, and reporting, as significant, the sets of genes associated to FDR lower than a threshold, typically 5%, or 1%, that in this case captures the expected fraction of features erroneously included in the sets of differentially expressed ones.

3.5. Enrichment analysis

Enrichment analysis is a common procedure in -omics pipelines to interpret the outputs of differential expression analysis, providing biological insight on the features of interest. These analyses typically take as an input a set of genomic features of interest (e.g. a set of differentially expressed genes) and asks whether the intersection between the set and another group of genes associated to an ontological label (e.g. genes associated to a given biological process, metabolic pathway, or signaling route, annotated as such in an external database, experiment, or source of knowledge) is larger, or smaller, than expected by chance. This is done by comparing the observed size of the intersection to a null-model-based statistical estimate of its expected size if elements in both sets were defined randomly, which allow us to compute significance values for each enrichment found.

To conduct these analyses, having access to exhaustive sets of external ontological annotations to confront to our sets of differentially expressed features is key. Two important examples of databases compiling extremely valuable sets of annotations are the Gene Ontology Consortium (GO) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) (24).

These two databases feature a hierarchical structure, where feature annotations are organized into hierarchical trees that may not be easy to interpret. This feature is incorporated to the analyses by programs like the app ClueGO from Cytoscape, which will be used in this work.

This software allows to search for biological paths in which a collection of genes are enriched with great flexibility in both the terms to search and the detail of the pathways to show. It compares the input genes to a collection of pathways with their relevant genes associated and studies the intersection of the input with the pathway members.

The output of ClueGO is a network where the nodes correspond to the external ontology labels tested whose intersection with the input set of features is significantly big (i.e. larger than expected by chance). The nodes are connected when the amount of features shared between the groups exceeds a threshold, providing a graphical and informative way of representing the biological functions and processes enriched among the genes of interest that act as input. Additionally, a table with relevant information about the network terms is provided along the graph itself.

ClueGO allows to implement enrichment analyses using several sources of external functional annotations. Arguably the most widely used is the Gene-Ontology Consortium (GO), that gathers extensive functional annotations, spanning from extremely generic to extremely specific ontology labels associated to biological processes, biochemical activity, and cell location.

KEGG, on the other hand, is a database that contains information about many biology related terms, ranging from information about specific genes to whole pathways. The aim of KEGG is to offer a comprehensive overview of all mechanisms involved in biological processes, relating pathways with genes, reactions, and connected pathways. An important utility of KEGG is that it offers curated and hand-drawn diagrams of common pathways in living organisms, with annotations on discovered proteins involved in each pathway in the organism of interest. This makes KEGG an useful resource for information about specific function of proteins.

4. Project Background and collaboration

This project has been carried on in the context of a collaboration with Dr. Rafael Prados Rosales (25), aiming to elucidate the mechanisms by which VirR regulates extracellular vesicle production. For this task, the two datasets that will be analyzed and interpreted in this Final Master Project were received from the partner research group, as well as further experimental context about the phenotypic differences between the different strains under analyses. This allowed for a more guided study of the obtained results, helping to better interpret the outputs of the experiments in the context of the phenotypic differences found.

In these experiments, the team characterized extensive changes in the cell wall structure of the VirR-KO strain. Cryo-electron microscopy revealed that VirR-KO cells had a thicker cell envelope caused mainly by an expansion of the cytoplasmic membrane, which suggested a deficient production of cell wall components present in the nearest part to the cytoplasmic membrane. After complementation with a functional copy of VirR (VirR-C), the structure of the cell wall returned to normal, giving more evidence to the role of VirR in maintaining a correct cell wall structure.

Another result that supported the hypothesis of a deficient cell wall after eliminating the functional VirR protein was the enhanced vesiculation of virR mutant M.tb. cells, confirmed by several experiments that suggested that cell envelope permeability directly correlates to increased production of extracellular vesicles.

Finally, lipidomic analyses were also conducted, pointing toward significant differences between WT and mutant strains when looking at phthiocerol dimycocerosate (PDIM) production, more abundant in mutant cells. PDIMs are important lipids in M.tb. which act as virulence factors and are related to the ability of M.tb. cells to escape defense responses of the host cells by disrupting their membranes (26).

These results, taking together support for a significant role for VirR in regulating cell wall composition, permeability, and function; aspects whose trace -we hypothesize- should be reflected into significant differences in transcriptomic and proteomic profiles.

5. Objectives

With all this information in mind, we can define the main objectives of this project:

- To determine the relation of VirR to M.tb. cell wall integrity, in order to check if regulation of cell wall processes was one of the mechanisms by which VirR was able to regulate EV expression.
- As the obtained datasets included information about many different genes and proteins, another goal was to take a closer look at overall VirR function in cell processes, not only on the cell wall ones, as that could give a clearer insight on VirR action mechanisms and relevance in M.tb. cells, either in basic cell functions or in virulence-related functions.
- Finally, another important goal of this work was to better understand extracellular vesicle function and the role of VirR in the regulation of its composition, by taking a comparative look at results of proteomics analysis between WT and VirR mutant cells to see if there were any differences in the protein cargo of the vesicles.

6. Materials and methods

6.1. Transcriptomic dataset description

The transcriptomic dataset used in this work consisted of RNA-seq raw count data obtained from an Illumina HiSeq2500 sequencer. The dataset contained counts of 4031 genes for seven different samples, capturing the number of RNA fragments mapping to each of the genes, in each sample, in a sequencing assay. Three of the samples corresponded to WT M.tb. cells, and four of them to VirR-KO mutant cells. These cells had been cultured using Iron (Fe)-depleted minimal medium (MM), with glycerol as the only carbon source.

As there were only two conditions on transcriptomics data, the experimental design consisted in a mono-factorial design using as factor the strain condition (See Figure 5).

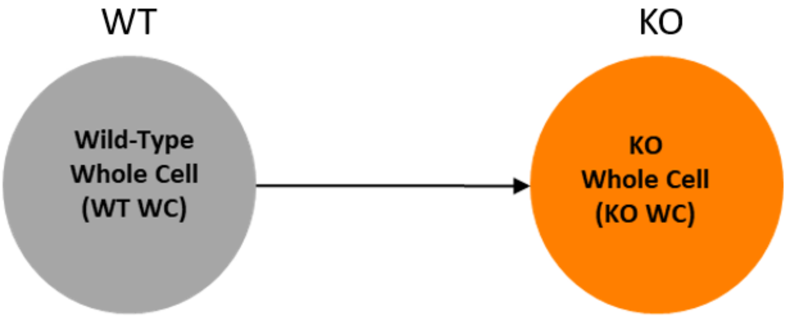


Figure 5. Experimental design of transcriptomics data

6.2. Proteomic datasets description

The proteomic datasets consisted of output data from a Label-Free Quantitative Mass Spectrometry experiment. The datasets had already been treated with Progenesis software (Nonlinear Dynamics) upon their reception and contained values for raw protein abundances.

A first set of samples corresponds to analysis on whole cell lysates. It includes quantitative data about 937 proteins, from 9 different samples, divided in groups of three replicates of each strain. The strains correspond to WT M.tb. cells, VirR-KO mutant cells, and VirR complemented M.tb. cells (VirR- C). These cells had been cultured in the same conditions as the cells from which the transcriptomic data was obtained.

The other sample set corresponds to proteomic profiles extracted from extracellular vesicles, containing data from 939 proteins, with the same number of samples grouped in the same way as the whole cell dataset.

Both sample sets were merged before analysis, studying only data of proteins found in both experiments in order to maintain comparability. The web service Db2db (27) was used to look for the corresponding official gene names for each gene, with the aim of facilitating the enrichment steps and identification of each item as the provided identifiers corresponded to GI numbers from GenBank.

The increased complexity of the proteomic data required a more complex design for its analysis. Thus, the proteomics experimental design is based on two factors: scope of the proteomics analysis (EVs or Whole cell) and correspondent strain of the samples (Wild Type, virR Knockout, virR Complemented), as shown in Figure 6.

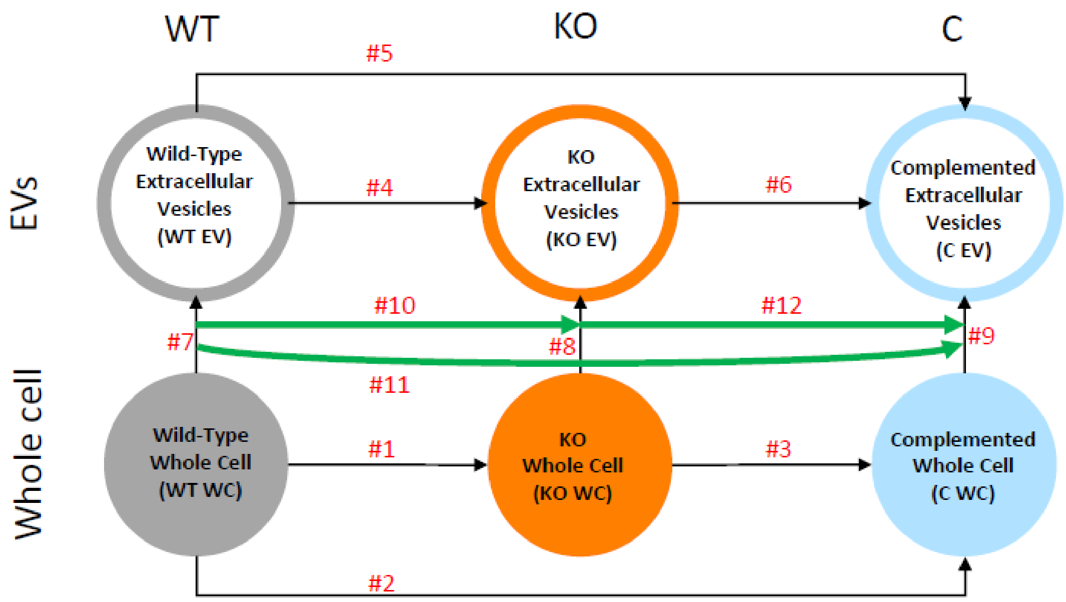

WT
KO
C
EVs
Whole cell
#5
Wild-Type Extracellular Vesicles (WT EV)
#4
KO Extracellular Vesicles (KO EV)
#6
Complemented Extracellular Vesicles (C EV)
#10
#12
#7
#8
#9
#11
Wild-Type Whole Cell (WT WC)
#1
KO Whole Cell (KO WC)
#3
Complemented Whole Cell (C WC)
#2

Figure 6. Experimental design for proteomic data

This implies a total of six conditions for the proteomics data, each one containing three biological replicates. Contrasts of first and second level were stablished between the different conditions, to check differences in related conditions and also interaction between different contrasts.

First level contrasts refer to direct comparison between two different comparable conditions, such as WT in WC versus KO in WC, which have in common the whole cell location and thus can be compared. This contrast would be written as follows:

$$(KO_{WC} - WT_{WC})$$

The first level contrasts of this analysis are contrasts #1 to #9.

Second level contrasts refer to differences between two first level contrasts, for example the difference between the WT vs KO contrast in the whole cell location and the same contrast in the EV location. This contrast would be written as follows:

$$(KO_{EV} - WT_{EV}) - (KO_{WC} - WT_{WC})$$

We can see that second level contrasts can be interpreted as the difference between two differences. These contrasts do not reveal differentially expressed genes between conditions but can reveal proteins which expression responds differently to a change in a certain condition depending on another condition. The second level contrasts of this design are contrasts #10, #11 and #12.

By establishing all these contrasts, we could extract as much information about our dataset as possible. The most important contrasts for our study are contrasts #1 and #4, which correspond to the comparisons between the KO and WT cells, either in WC or in EV. The KO vs WT at WC contrast is important because, within this setup, it offers the more analogous contrast to the differences KO-WT that will be estimated from transcriptomic data. In the case of the KO vs WT at EV contrast, the importance resides in the fact that it can reveal the differences in the cargo of the M.tb. EVs when the function of VirR is absent, which is crucial for determining whether VirR has any implication on its regulation or not.

Regarding the other contrasts, #3 and #6 give information about the restoration of functions when complementing the KO bacteria with a functional copy of VirR. #2 and #6 allow to check if the complemented phenotype is similar to the WT one. #7, #8 and #9 can show the differences in protein cargo between whole cells and EVs. Finally, contrasts #10, #11 and #12 can reveal differences in expression when changing the condition and location.

6.3. Transcriptomics pipeline

6.3.1. Preliminary assessment

PCA was performed on the transcriptomic samples with the goal of having a preliminary overview of the characteristics of the dataset and the relation between the samples. For the PCA analysis, normalized reads from Deseq2 (28) were used. The method that Deseq2 uses to normalize the counts is explained in the next section.

6.3.2. Differential expression analysis

Differential expression analysis of transcriptomic data was made using the R package Deseq2 (28), available from Bioconductor. The raw counts data from an RNAseq experiment (our input) is not suitable for differential expression analysis, and it is necessary to transform the data into a variable that can be compared across samples and genes. Deseq2 achieves this goal by calculating the geometric mean for each gene across all samples. Then, the counts for each gene in each sample are divided by that mean:

$$R'^{j}_{A} = \frac{R_{raw}{}^{j}_{A}}{\sqrt[j]{R_{raw}{}^{1}_{A} * R_{raw}{}^{2}_{A} * \ldots * R_{raw}{}^{j}_{A}}}$$

Where $R_{raw}{}^{j}_{A}$ represents the raw read count of gene A in sample j, and R'^{j}_{A} represents the corresponding transformed value.

Finally, it uses the median of those new values as a size factor for each sample. This size factor is a value that is used to normalize the counts, by dividing each count for its corresponding size factor. Then, it models the normalized counts obtained this way assuming that the raw count data are distributed according to a negative binomial distribution.

$$R^{j}_{A} \sim NB(\mu, \alpha)$$

Where R^{j}_{A} represents the normalized read count of gene A in sample j, μ is the expected value of the expression, and α represents the amount of overdispersion of the variable, which is related to biological variability.

Deseq2 fits a generalized linear model based on this distribution, using the transformed data to estimate the expected value of the distribution. As for the overdispersion parameter, Deseq2 initially estimates it for each gene using Maximum Likelihood Estimation, and then applies a correction method called Bayesian shrinkage to produce maximum a posteriori estimates using the MLE combined with priors that are defined by integrating the data on all genomic features analyzed. Finally, after these computations, Deseq2 computes each genes' logarithmic Fold Changes (logFCs, also called effect sizes) for each specified contrast, extracts standard errors from the models which end up providing p-values.

Resulting p-values from the contrast were treated for multiple testing error using *fdrtool*. Genes with False-discovery rates lower than 0.05 were marked as significant. The specific algorithm used for multiple testing correction is the Storey-Tibshirani method, which models the distribution of p-values as a mixture of true positives (enriched in low values), and null hypothesis (modelled as a flat p-value histogram). By estimating the expected proportion of the latter, denoted as π_o, the false discovey rate associated to each p-value can be easily estimated as this: $fdr(p *) = \pi_o \cdot p */F(p *)$, where $F(p)$ is the fraction of genetic features with a p value lower than p*. For transcriptomics, we labelled as significant all tests with FDR<0.05. Finally, the sign of the LogFCs was used to determine upregulation or downregulation in the studied strain with respect to the reference.

6.4. Proteomic pipeline

6.4.1. Normalization

Among the assorted group of software packages available in R gathering methods for the completion of the different steps needed for the analysis of label-free proteomic data, such as DEP (29), msqRob2 (30–32) and WrProteo (33), the selected method for normalization of the proteomic data was the built-in method for the software package *DEP (29)*, which is *vsn* (34), available as a package in Bioconductor. *vsn*, short for "Variance Stabilizing Normalization", is a method originally designed to be used with microarray data for gene expression analysis, but that can be applied to other input types. *vsn* is to correct heteroskedasticity patterns through which the variance of the intensity of a certain measurement, being it the intensity of a spot in a microarray or in a LF-MS proteomics dataset like in our case, varies along with the mean intensity measured for that specific feature across samples. This issue generates a difficulty when trying to apply linear models to compare expression between different samples, since the vast majority of those operate under the assumption of homoskedasticity, that is, that the variance is homogeneous independently from the mean intensity of the signal. *vsn* aims to overcome this situation by transforming the data in a way that eliminates the dependence of the variance with respect to the mean.

6.4.3. Imputation of missing values

For imputing the missing values in the datasets, the chosen method was Bayesian Principal Component Analysis (*bpca*) (35), which is available as one of the several imputation methods that are present in the *DEP* package.

bpca was chosen after a bibliographic review of benchmarking of imputation methods. This method has been reported to offer a good performance for global imputation when all types of missing values coexist in a dataset (36). The method was mainly developed for imputing MAR values but is also able to handle MNAR values too.

bpca is based on a principal component regression, which consists in performing a PCA considering only data with no missing values and then, after obtaining the descriptors of the analysis, estimating the missing values based on the observed part by using the result of the PCA. The method combines a probabilistic approach for the PC regression, and a repetitive algorithm that iterates through estimated missing values in order to refine the results, until reaching convergence.

6.4.4. Preliminary assessment

PCA was also performed on proteomics data before differential expression analysis. In this case, we performed separate PCAs on the samples depending on their location, and also a global PCA of all proteomics samples at the same time. PCA was performed after normalization and imputation.

6.4.5. Differential expression analysis of proteomics data

Differential expression analysis was made by using the *limma* (37) package for R.

The package *limma* was originally developed as a method to analyze microarray data, but with several adaptations can be used for both proteomics or RNA-seq data. As we had already normalized and imputed for missing values, direct modeling with *limma* was possible. *limma* uses linear models to estimate the logFCs of each gene for each defined contrast. Then, it computes the standard errors and applies a Bayesian shrinkage procedure to refine the significance of the tests.

The resulting p-values from the *limma* modelling step for the significance of the expression of each protein, for each contrast, were treated for multiple test correction using again *fdrtool* (38) In this case, considering the lower statistical power of the tests, proteins with an FDR<0.1 were marked as significant for differential expression regarding the corresponding comparison. The sign of the resulting LogFCs was used as an indicator of upregulation or downregulation of genes in each condition with respect to the corresponding reference for each contrast.

6.5. Enrichment of proteomics and transcriptomics data

Pathway enrichment of the differentially expressed proteins and genes was made using the ClueGO (39) app from Cytoscape (40). A network and its corresponding table was retrieved for each contrast and direction of the DE features, i.e. each of the contrasts has a network related to upregulated pathways and another one for downregulated ones. Only contrasts with enough terms to obtain a significant network after separating between upregulated and downregulated ones were selected for discussion in this work. Enrichment terms from Gene Ontology (41,42) for the biological process and cellular component of both proteomics and transcriptomics output data were searched, for each contrast.

The GO term Fusion feature was selected in order to reduce redundancy in the enrichment output data. The FDR threshold for significant pathways was selected at 0.1. The Benjamini-Hochberg method was used for correction of p-values.

An interval between 4 and 10 for GO hierarchical levels was selected, to focus the search on intermediate level terms, neither too generic, nor too specific. For the proteomic enrichment, clusters with at least 1 gene and for which the matching genes in the input set represented at least 15% of the total genes of the enrichment cluster were selected. In the case of transcriptomic data, the selection criteria for the clusters were to contain at least 10 genes from the list, which had to represent at least 25% of the total number of genes of the cluster.

6.6. Integration of proteomic and transcriptomic results

Integration of both -omics techniques was explored by selecting only the genes and associated encoded proteins that are shared between both the transcriptomics and proteomics datasets, and then checking the coincidence between the logFC direction and size for each set. Genes and proteins fulfilling statistical significance criteria in both cases were divided according to the sign of their logFC, marking those with a positive value as upregulated in KO condition and those with a negative value as downregulated in KO condition, for both datasets. A Fisher Exact Test was performed on the data to check for the statistical enrichment in the intersection of significant effects between transcriptomics and proteomics. The resulting odds-ratio and associated p-value was used to determine that relevance.

7. Results and discussion

7.1. Principal component analysis and differential expression analysis

A PCA was performed on transcriptomic data to check differences between samples prior to quantitative analysis.

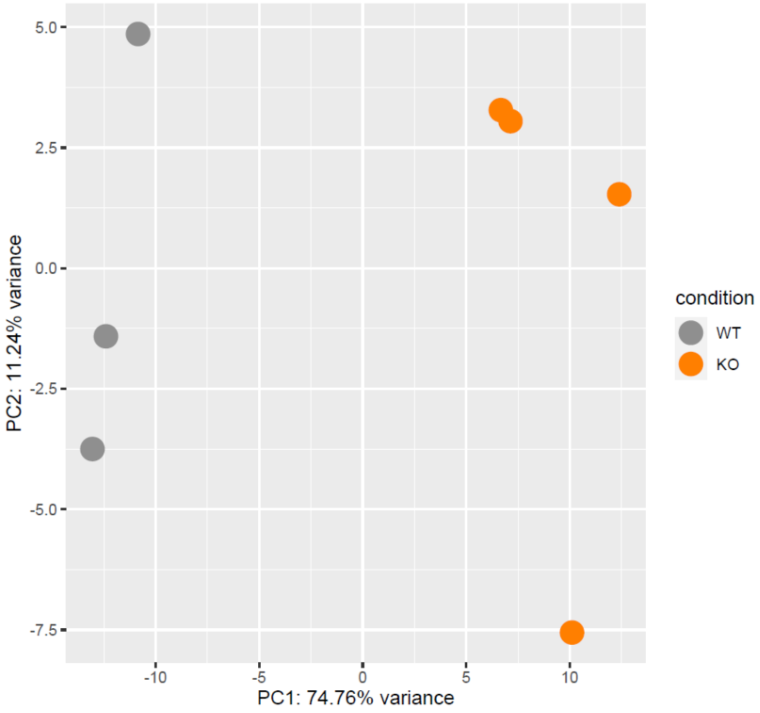


Figure 7. PCA of transcriptomic data. The two principal components with a higher explanation of the variance were used for the plot. In grey, the samples corresponding to the WT condition. In orange, the samples corresponding to the KO condition.

The PCA plot reveals that there are very strong differences between samples of KO and WT conditions. The first principal component, which explains 71.99% of the variance of the dataset, clearly separates samples between conditions (r= 0.9846568, p= 5.555e-05). Biological replicates of both strains, although very dispersed in the graph, are close in the x axis, leaving almost all of the separation for the second principal component, which only explains 11.95% of the variance. This shows that differences between samples within strain are smaller than differences across.

This plot suggests that absence of a functional copy of VirR leads to a completely different transcriptomic landscape. To validate that hypothesis on a per-gene basis, we conducted DEA, as described in the methods section. The results of these analyses can be summarized in the following volcano plot:

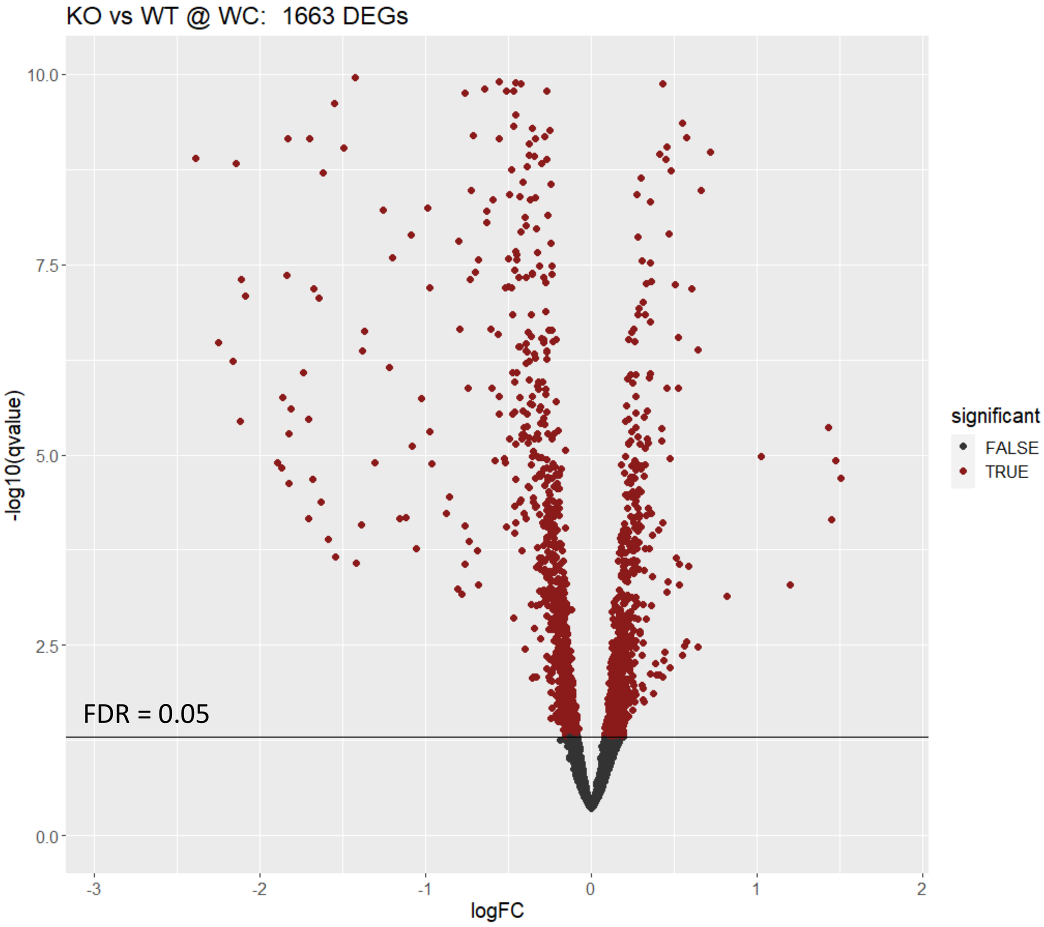


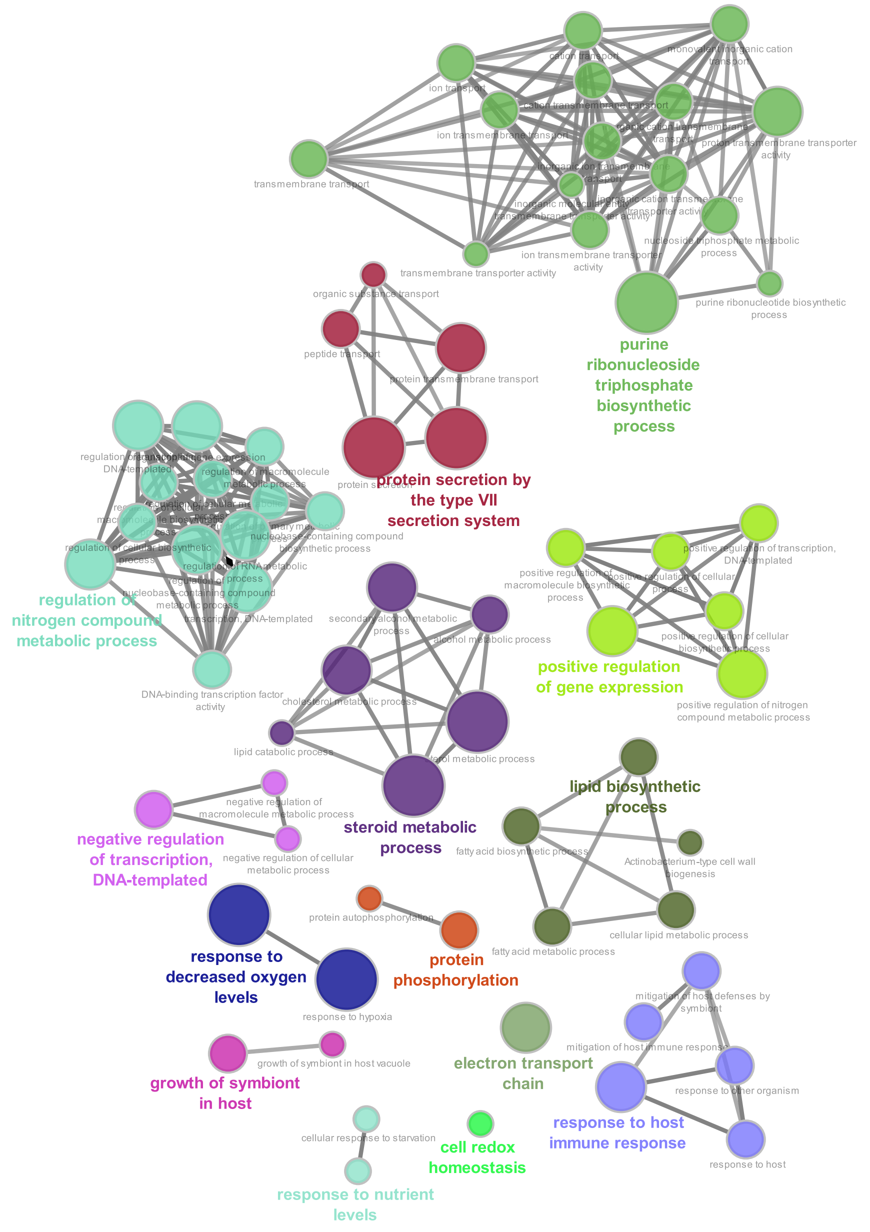
Figure 8. Volcano plot of the differentially expressed genes found in transcriptomic data

The volcano plot shows at the same time the intensity of the difference between the expression of a protein in one condition and in another, represented by the x axis (logFC), and the significance of said difference in the y axis (-log of the corrected p-value). Genes at the left side were upregulated in the WT condition, and genes at the right side were upregulated in the KO condition. The number of differentially expressed genes found in transcriptomic data was 1663. 748 of these genes were upregulated in KO mutant cells, and 915 were upregulated in WT cells. This represents a high proportion of the total genome of M.tb. (More than 22% of the genome), which highlights the relevance of VirR as a systemic regulator of gene expression.

7.2. Transcriptomics Enrichment Networks

The large sets of upregulated and downregulated differentially expressed genes allowed for a separate enrichment analysis using ClueGO stratified by the direction of the effects.

Downregulated genes in KO condition



cation transport
monovalent inorganic cation transport
ion transport
cation transmembrane transport
ion transmembrane transport
inorganic cation transmembrane transport
proton transmembrane transporter activity
transmembrane transport
inorganic ion transmembrane transport
inorganic cation transmembrane transporter activity
inorganic molecular entity transmembrane transporter activity
nucleoside triphosphate metabolic process
ion transmembrane transporter activity
transmembrane transporter activity
purine ribonucleotide biosynthetic process
purine ribonucleoside triphosphate biosynthetic process
organic substance transport
peptide transport
protein transmembrane transport
protein secretion
protein secretion by the type VII secretion system
regulation of transcription, DNA-templated
regulation of gene expression
regulation of macromolecule metabolic process
regulation of cellular metabolic process
regulation of cellular macromolecule biosynthetic process
regulation of primary metabolic process
nucleobase-containing compound biosynthetic process
regulation of cellular biosynthetic process
regulation of RNA metabolic process
regulation of nucleobase-containing compound metabolic process
transcription, DNA-templated
regulation of nitrogen compound metabolic process
DNA-binding transcription factor activity
secondary alcohol metabolic process
alcohol metabolic process
cholesterol metabolic process
lipid catabolic process
sterol metabolic process
steroid metabolic process
positive regulation of transcription, DNA-templated
positive regulation of macromolecule biosynthetic process
positive regulation of cellular metabolic process
positive regulation of cellular biosynthetic process
positive regulation of gene expression
positive regulation of nitrogen compound metabolic process
negative regulation of macromolecule metabolic process
negative regulation of transcription, DNA-templated
negative regulation of cellular metabolic process
lipid biosynthetic process
fatty acid biosynthetic process
Actinobacterium-type cell wall biogenesis
cellular lipid metabolic process
fatty acid metabolic process
protein autophosphorylation
protein phosphorylation
response to decreased oxygen levels
response to hypoxia
growth of symbiont in host vacuole
growth of symbiont in host
electron transport chain
mitigation of host defenses by symbiont
mitigation of host immune response
response to other organism
response to host immune response
response to host
cellular response to starvation
cell redox homeostasis
response to nutrient levels

Figure 9. Enrichment network of the differentially expressed genes downregulated in the KO condition.

The enrichment analysis of the downregulated terms in the VirR-KO strain revealed significant enrichments in terms related to response to stressful conditions, such as starvation (FDR=7,88e-2) or hypoxia (FDR=6,67e-8), and also related to interaction with host processes, like its immune response (FDR=2,07e-3). Many terms related to regulation of gene expression were also found (e.g. positive regulation of gene expression, FDR=1,91e-3), as well as terms related to protein secretion (e.g. protein secretion by the type VII secretion system, FDR=1,01e-5). These results are coherent with previous knowledge of the decreased virulence of VirR mutant M.tb. strains.

Upregulated genes in KO condition

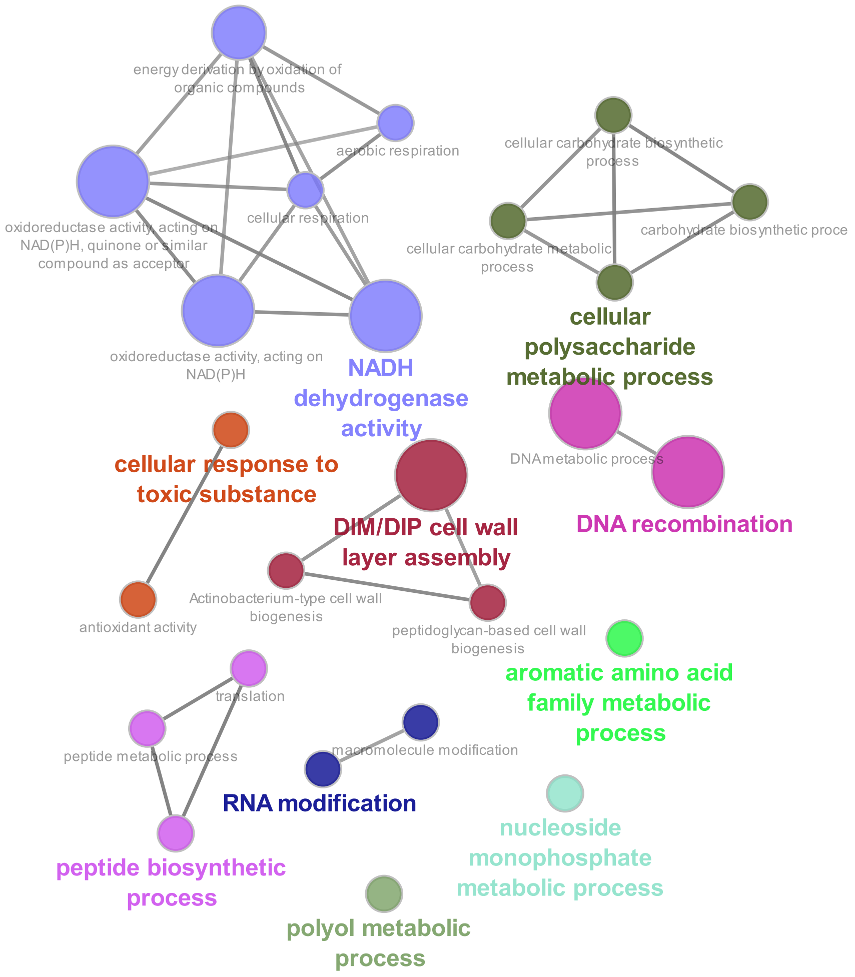

energy derivation by oxidation of organic compounds
aerobic respiration
cellular respiration
oxidoreductase activity, acting on NAD(P)H, quinone or similar compound as acceptor
oxidoreductase activity, acting on NAD(P)H
NADH dehydrogenase activity
cellular carbohydrate biosynthetic process
carbohydrate biosynthetic proce
cellular carbohydrate metabolic process
cellular polysaccharide metabolic process
DNAmetabolic process
DNA recombination
cellular response to toxic substance
antioxidant activity
DIM/DIP cell wall layer assembly
Actinobacterium-type cell wall biogenesis
peptidoglycan-based cell wall biogenesis
aromatic amino acid family metabolic process
translation
peptide metabolic process
peptide biosynthetic process
macromolecule modification
RNA modification
nucleoside monophosphate metabolic process
polyol metabolic process

Figure 10. Enrichment network of the differentially expressed genes upregulated in the KO condition.

Regarding the upregulated genes in the KO condition, we found terms related to response to toxic substances (FDR=8,21e-2). These results, along with the prominent enrichment among ion transport (FDR=2,84e-2), and related terms, among genes more expressed in the WT strain are coherent with the increased permeability found in the VirR-KO using electron microscopy, which would enable the bacterium to disregard the transcriptional machinery that enables transport of metallic ions, at the same time it requires from the bacterium a detoxification effort.

Furthermore, we also see how metabolism is widely reprogrammed in the mutant strain. On the one hand, carbohydrate biosynthesis related terms (FDR=7,67e-2) were enriched among genes that are upregulated in the mutant, as well as terms related to NADH dehydrogenase activity (FDR =5,37e-4). On the other hand, terms related to metabolism of steroids (FDR=9,21e-5) and lipid biosynthesis (FDR=6,13e-3) were enriched among genes down-regulated in the VirR_KO strain. Finally, certain highly important terms that were found enriched among genes more expressed in the mutant strain were related to biogenesis of the cell wall and synthesis of phthiocerol dimycocerosates (DIM/DIP cell wall layer assembly). These molecules, shortened as PDIMs, are a type of virulence factors (43) that are usually produced by M.tb. as a way to detoxify propionate, a common product of catabolism of cholesterol and lipids. This way, M.tb. cells are able to lose a potential toxic agent while at the same time increasing their pathogenic nature towards the host. The original cells were cultured using glycerol as the only carbon source, and this glycerol can be converted into propionyl-coA, which is a byproduct of propionate metabolism (44).

Importantly, not all the enrichment terms related to metabolic pathways that were found were equally easy to interpret, and some of them were arguably too general to provide useful description of the processes that are altered between both strains, especially the terms related to cellular polysaccharide metabolic process and NADH dehydrogenase activity. Because of that, we selected the genes that were causing that enrichment, searched for their activities and characteristics in KEGG and Uniprot (45), and produced the following heatmaps to check their expression value:

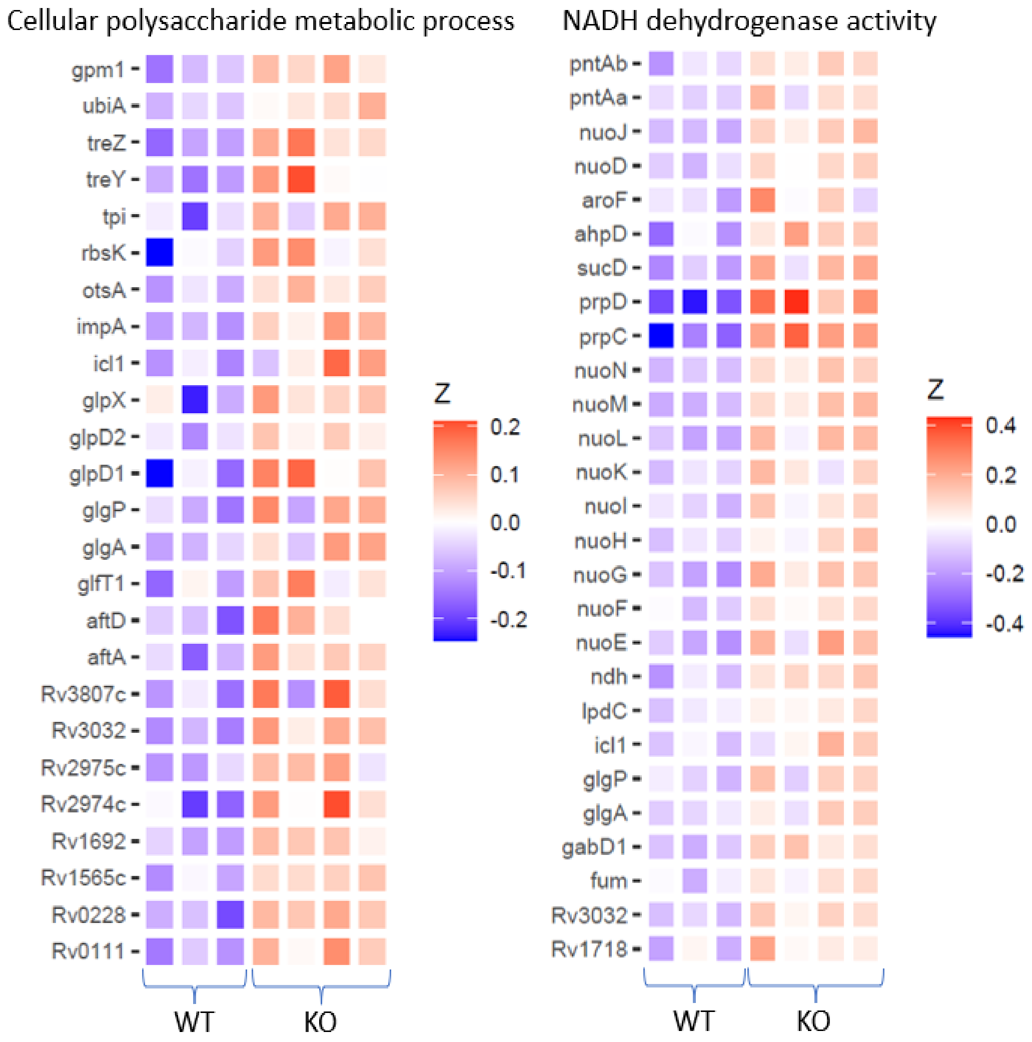


Figure 11. Left: Heatmap of genes related to enrichment in terms related to cellular polysaccharide metabolic process. Right: Heatmap of genes related to NADH dehydrogenase activity. A higher Z score corresponds to a higher relative expression.

Analysis of the genes related to the cellular polysaccharide metabolic process revealed that many of these genes were related to cell wall processes. Among these genes, we found genes related to trehalose synthesis and metabolism (otsA, treZ, treY), myo-inositol synthesis (impA), propionate degradation (icl1) and cell wall polysaccharide synthesis (glgA, glfT1, aftD, aftA Rv3807c, Rv3032). These terms suggest an overexpression of cell wall components that can be linked to the enlarged cell wall observed in electron microscopy. Several of these genes are related to arabinogalactan synthesis (glfT1, aftD, aftA), supporting the hypothesis of an aberrant production of components of the peptide wall that are close to the membrane sketched in Figure 2.

On the other hand, analysis of the genes related to NADH dehydrogenase activity revealed that some of the genes were related to propionate degradation (prpD, prpC, icl1) while most of them were related to aerobic respiration and citric acid cycle (nuo group, ndh, sucD). Some common genes with the polysaccharide metabolism related ones were found too (icl1, glgA,glgP, rv3032).

The presence of genes involved in propionate detoxification is coherent with the upregulation of PDIM synthesis related terms. According to these results, we can hypothesize that M.tb. is increasing PDIM production to detoxify the propionyl-coA that is produced as a by-product of glycerol metabolism, even if our data delineate an intriguing, uncommon outlook wherein increased production of PDIM, in our case appears associated to lower virulence levels in the VirR-KO strain, suggesting, perhaps, a compensatory adaptation mechanism by the bacteria.

Finally, we also observe a number of terms enriched among genes more expressed in the VirR_KO than in the WT, that are related to translation and peptide biosynthesis (FDR=5,76e-2), as well as terms related to RNA modification (FDR=5,61e-2). These results suggest a differential involvement in post-transcriptional regulatory mechanisms in the mutant strain, which perhaps may cause proteomic profiles nbot fully recapitulated by the transcriptomic data. This results further motivates inspection of proteomic profiles in WT and VirR-KO strains to fully understand the regulatory role of the virR protein and its involvement in modulating M.tb. permeability, virulence, metabolism; cell-wall composition and vesiculation.

7.3. Proteomics Principal Component Analyses

Before the differential expression analysis on the proteomics data, a PCA was conducted to check global differences in protein expression patterns across strains (WT, VirR-KO and, in this case, VirR-C) and locations (Whole cell -WC- and Extracellular vesicles alone -EV-).

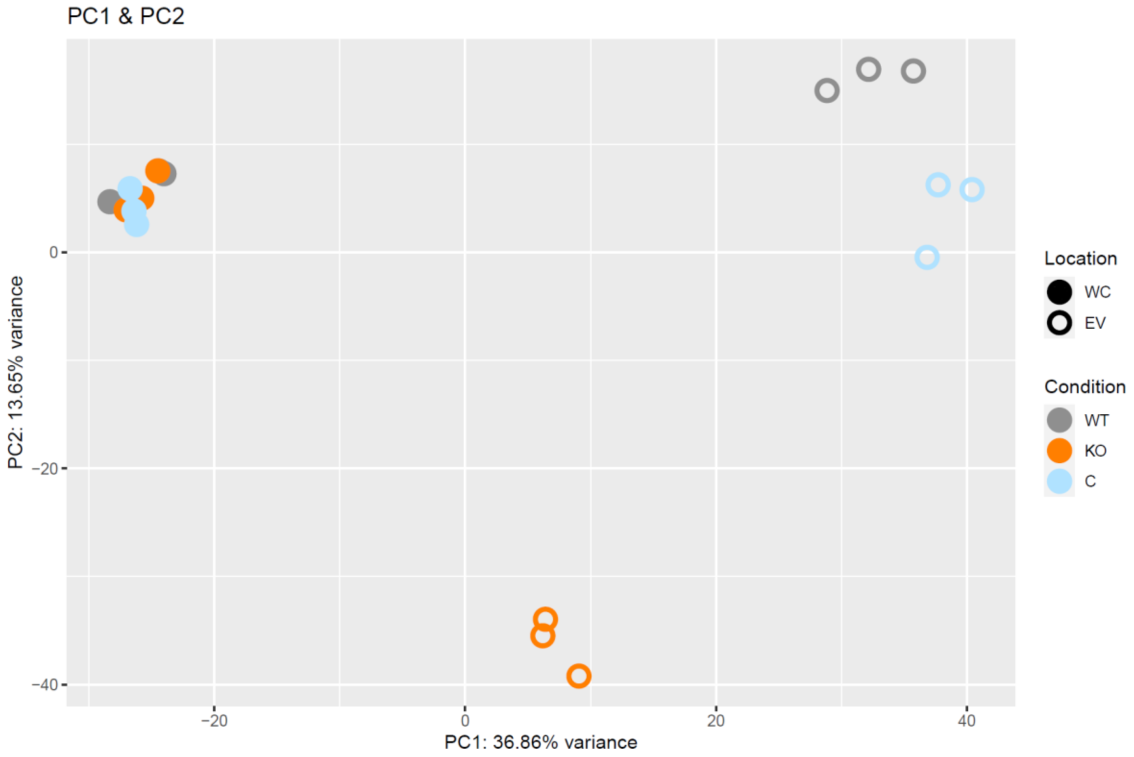


Figure 12. PCA of all samples

This PCA, summarized in figure 12, unveils some important facts about our data. First of all, we can see an important separation by the first principal component, which explains up to 36.9% of the variance of the original data, between samples from whole cell lysates and from isolated extracellular vesicles (r= 0.9285992, p= 7.532e-08), indicating that cellular compartment is indeed the most decisive factor impacting proteomic profiles in our data. The second principal component correlates to presence or absence of virR regulation only in vesicles (r= 0.9646222, p= 4.32e-10), as demonstrated by the position of the KO EV samples only. Moreover, although complementation of VirR does not restore the original phenotype in EV samples, the new phenotype, associated to an increased protein expression of VirR (see figure 13), is clearly closer to WT strains than to the KO ones.

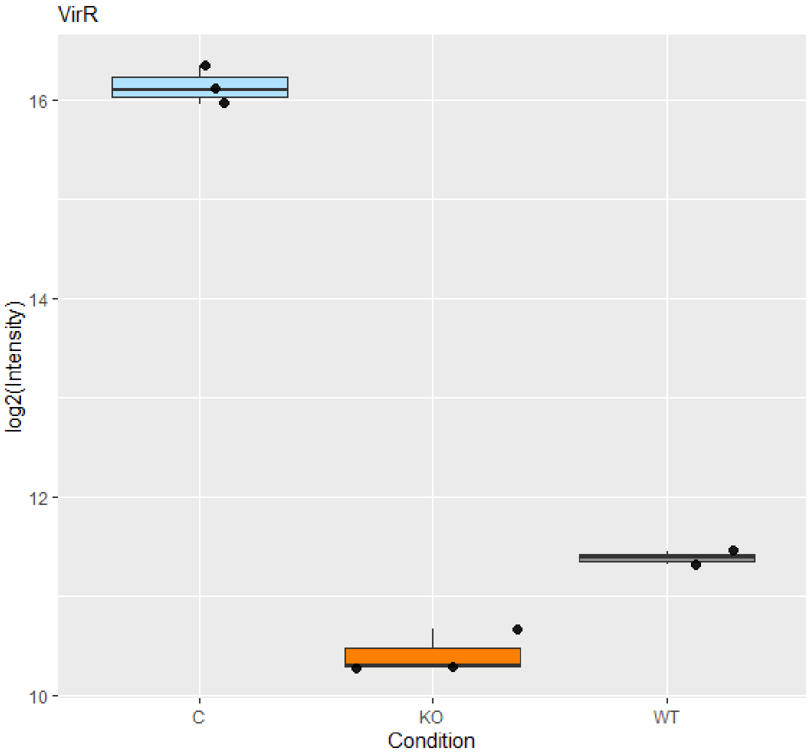


Figure 12. Expression plot of VirR in WC lysates. On y axis, the log normalized intensity of the recorded signal. On the x axis, the three different conditions of the samples.

In fact, analysis of the presence of virR in the samples is instrumental to interpret the proteome-wide differences found between strains. This analysis revealed an overexpression of the protein in the complemented strain, as well as presence in both the WT and KO condition. The overexpression of the protein in the complemented strain explains the differences between the WT and C samples, as *a priori* they should behave the same way, but an increase of the presence of a protein that is so relevant in regulation such as VirR is compatible with a change of the profile of the studied cells. The expression of VirR in the KO condition is explained by the fact that the VirR mutant did not have a deletion of the gene, but an insertion of a sequence that, when translated, creates a non-functional mutant of the protein.

Furthermore, we see that the differences between the proteomic profiles between strains in whole cell data, are residual when compared to either differences in cellular location or between strains in extracellular vesicles. Overall, this analysis points to the relevant conclusion that VirR effects on proteomic profiles focus primarily to the protein cargo of EVs, rather than to global whole-cell protein levels, which, in spite of the large transcriptomic differences found between strains, are relatively stable in spite to the presence or absence of the VirR regulator.

7.5. Proteomics Differential Expression Analysis

Following these analyses, the differential expression analysis was performed to take a quantitative look at the differences seen between conditions. The results of the analyses can be found in table 1 (See figure 6 for reference of the contrasts).

Nº	Contrast	Condition	Upregulated proteins in condition	Total DEP in contrast
#1	KO vs WT @ WC	KO	22	29
		WT	7	
#2	C vs WT @ WC	C	36	46
		WT	10	
#3	KO vs C @ WC	KO	18	40
		C	22	
#4	KO vs WT @ EV	KO	28	84
		WT	56	
#5	C vs WT @ EV	C	19	36
		WT	17	
#6	KO vs C @ EV	KO	10	40
		C	30	
#7	EV vs WC @ WT	EV	44	60
		WC	16	
#8	EV vs WC @ KO	EV	17	33
		WC	16	
#9	EV vs WC @ C	EV	35	55
		WC	20	
#10	Interaction @ KO vs WT	KO	7	26
		WT	19	
#11	Interaction @ C vs WT	C	1	15
		WT	14	
#12	Interaction @ KO vs C	KO	8	26
		C	18	

Table 1. Summary of the differentially expressed proteins found in each contrast, for each condition.

The contrast that had larger differences between the studied conditions was the comparison (KO vs WT) evaluated at EVs (84 DEPs at 10% FDR). In the case of the KO vs WT contrast at whole cell lysates (WC), only N=29 differentially expressed proteins were found, which suggests that VirR has a greater impact on the regulation of the protein identity of extracellular vesicles than in the regulation of the proteome of the whole cell.

To better understand the relations between the differentially expressed proteins found in the different contrasts, several Euler plots were produced, as shown below starting with the WC contrasts:

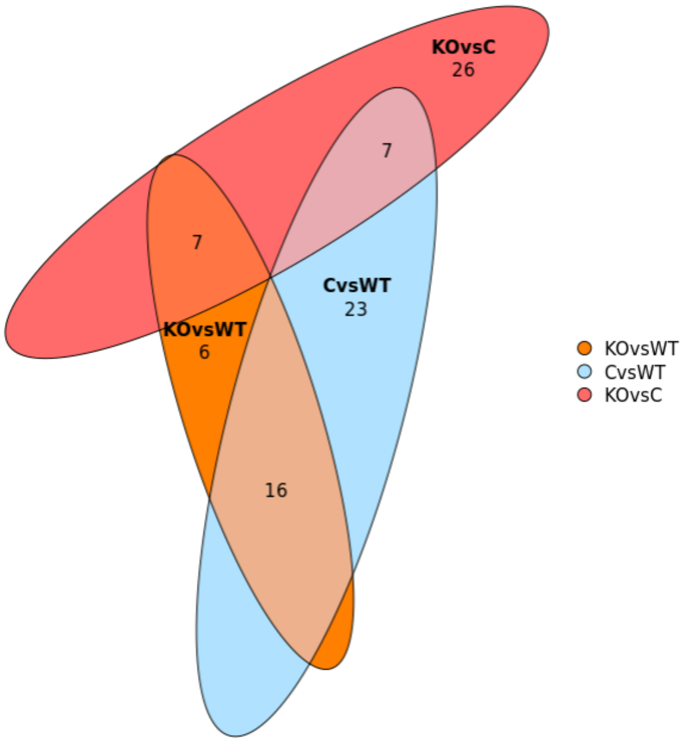


Figure 13. Euler plot of differentially expressed proteins found in WC contrasts

The WC contrasts overall revealed low numbers of differentially expressed proteins, which was expected after the PCAs. The intersections between the identities of the differentially expressed proteins found in the contrasts also correlated with what was observed in the PCA analysis, when looking at the position of the samples.

The next contrasts that were studied were the ones corresponding to EV location:

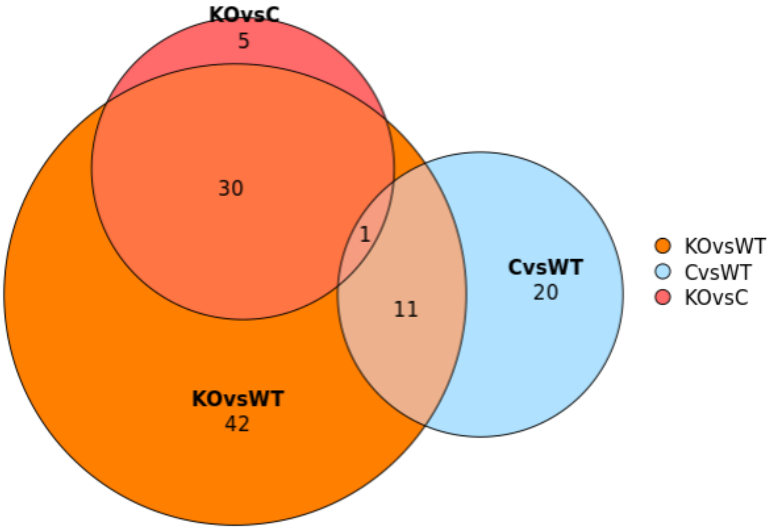


Figure 14. Euler plot of differentially expressed proteins found in EV contrasts.

There were more differentially expressed proteins found in EV contrasts than in those corresponding, supporting the hypothesis of the main role of EV cargo regulation by virR. Almost all of the proteins that were differentially expressed between the complemented and the KO conditions were also differentially expressed between the KO and WT ones, suggesting that the presence or absence of VirR is what mainly determines the proteomic profile of EVs.

Finally, the last Euler plot that was produced was the one corresponding to the EVs vs WC contrasts for each condition.

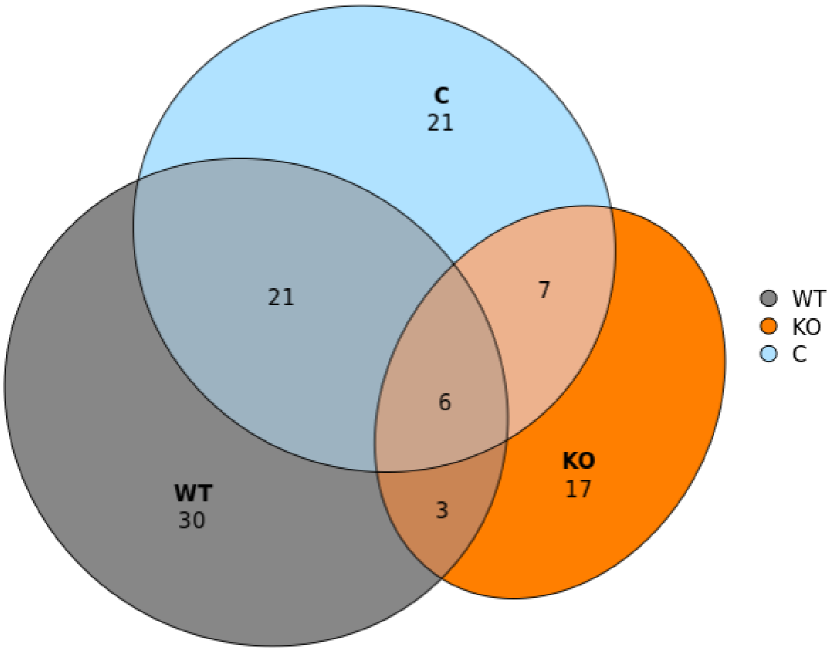


Figure 15. Euler plot of differentially expressed proteins found in WC vs EV contrasts.

The differences in the proteomic profiles between both locations on each condition had some similarities between the studied strains, which could mean that there are some proteins that are not affected by the function of VirR when setting their location. Nonetheless, these results confirm that there are important differences between the proteomic profiles of EV and their corresponding cells. The differences were notably smaller in the case of the KO condition.

From now on, we will focus on the main contrasts of interest in this study, which are the ones corresponding to the comparison between KO and WT conditions, both in whole cell lysates and in extracellular vesicles. Further analyses of the results of the analysis of the rest of contrasts can be found in the appendices.

To have a better look at the characteristics of the differential expression in each comparison, we produced the following volcano plots:

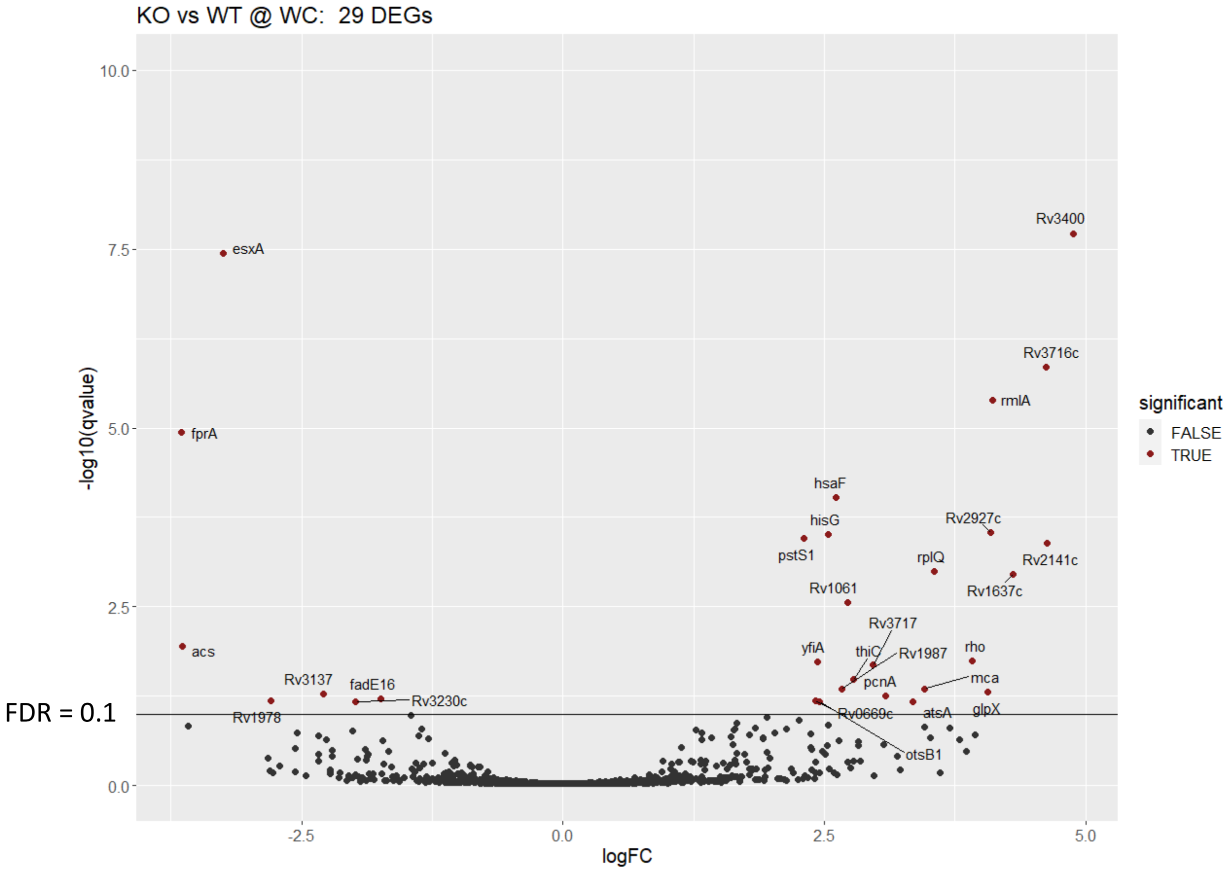


Figure 16. Volcano plot of the KO vs WT contrast in whole cell lysate samples

As seen before, there were few differentially expressed proteins in the KO vs WT @ WC contrast, with some of them having high differences in their expression. 29 differentially expressed proteins were found in the WC case, most of them being upregulated in the KO condition.

Next, we created another volcano plot, for the same contrast at the EV location:

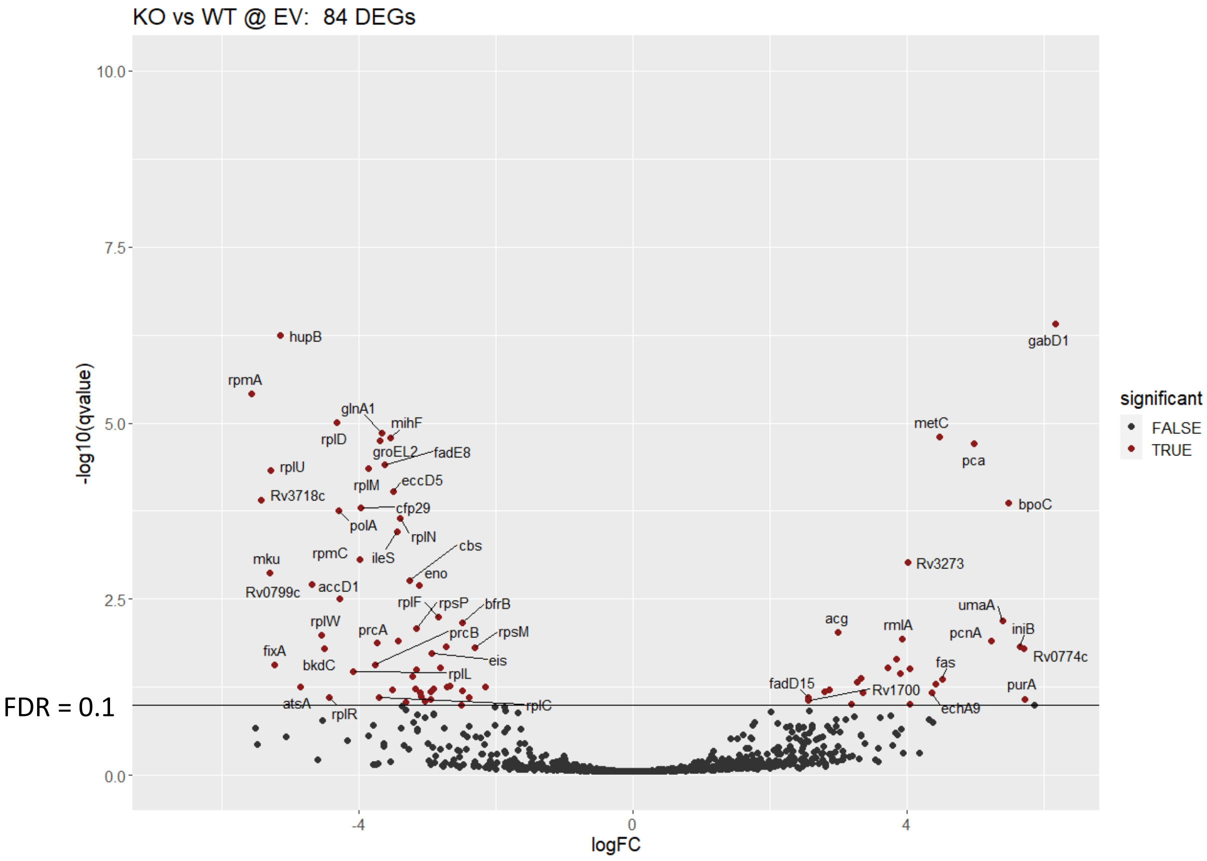


Figure 17. Volcano plot of the KO vs WT contrast in extracellular vesicles samples

In the EV case, 84 differentially expressed proteins were found. Here most of them were downregulated in the KO condition, which hints at a probable loss of function for the EVs as they seem to lose their original proteomic profile, possibility that we later evaluated.

We checked the concordance between the sets of differentially expressed proteins in both contrasts, to see if the absence of virR affected the same proteins at both locations. The following Euler plot was produced to confirm this idea:

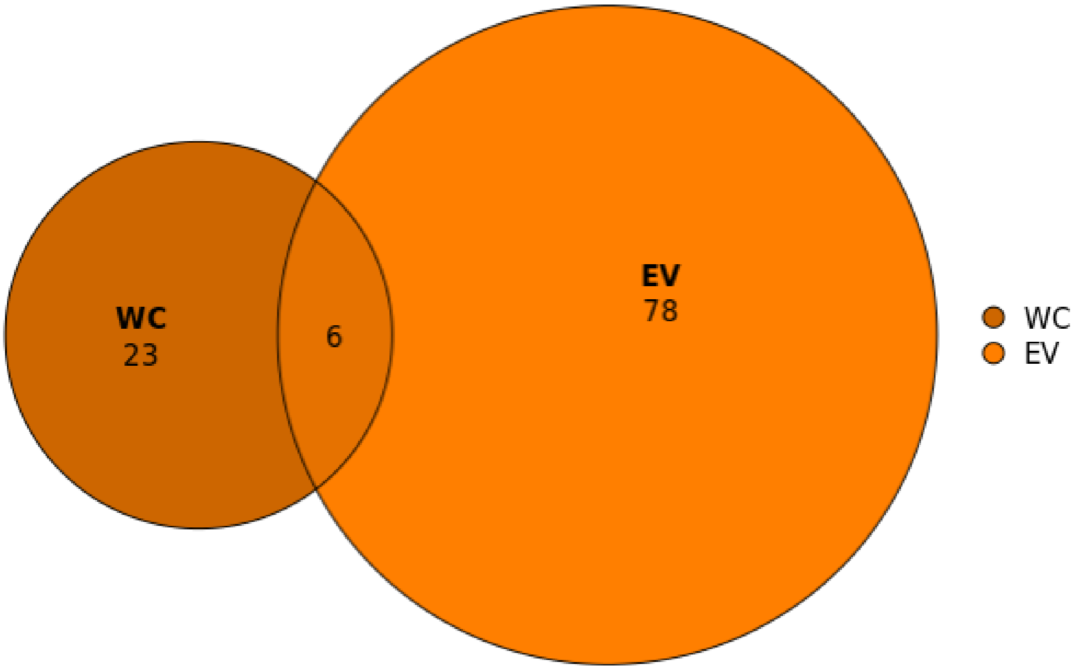


Figure 18. Euler plot of the coincident DE proteins in WC and EV samples, for the KO vs WT contrast

The Euler plot revealed that there was little coincidence between the differentially expressed proteins found in each of the two contrasts, supporting the idea of a large difference in regulation by virR depending on cell location.

7.6. Enrichment of proteomics differential expression analysis results

From the 29 differentially expressed genes in the KO vs WT contrast, 22 were upregulated and 7 downregulated for the KO condition. The low amount of differentially expressed proteins in the whole cell contrast did not allow for a separate enrichment analysis using ClueGO, as no significance was found in the associated terms for each gene. Because of that, we produced a heatmap to check the distribution of the expression values of each gene in the WT and KO conditions of the WC samples. The heatmap is presented below, with the associated function of each gene, as found in Uniprot (45):

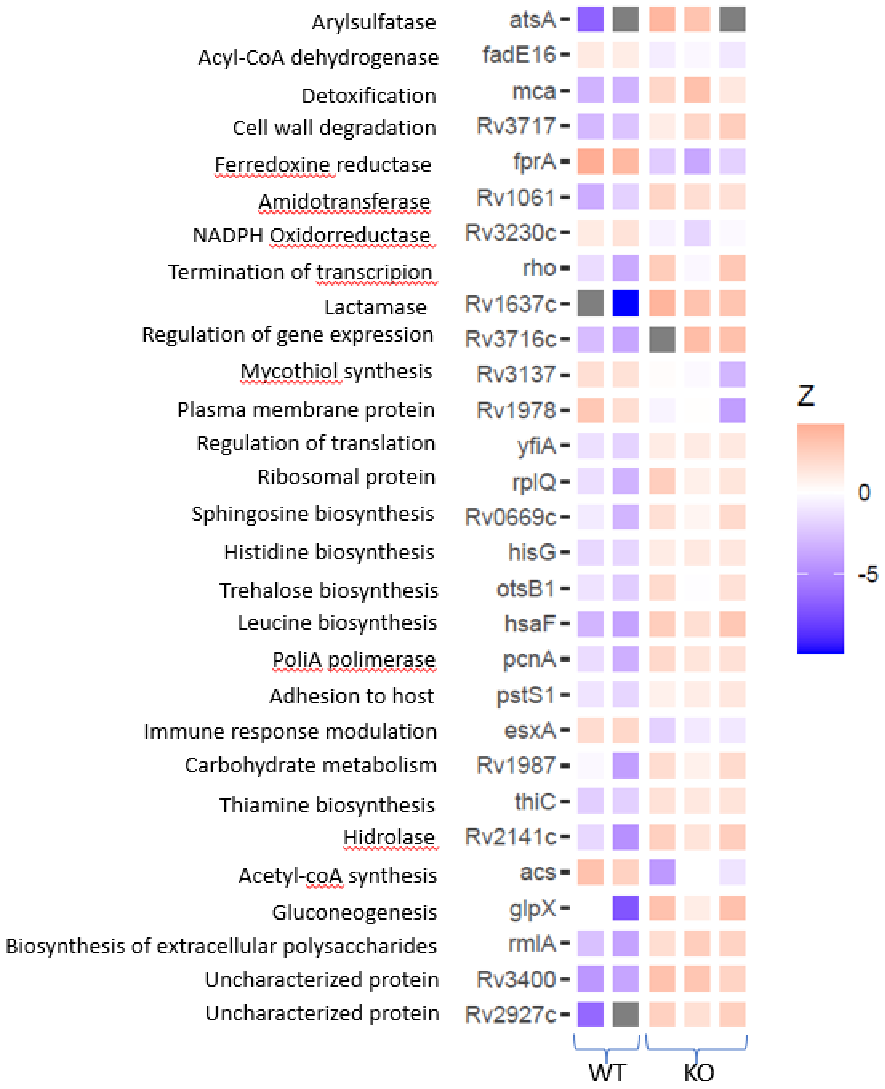


Figure 19. Heatmap of the differentially expressed genes between WT and KO conditions. A higher Z score corresponds to a higher relative expression. Grey squares represent missing values in raw data.

Among the genes that were downregulated in the mutant strain, we find several metabolism related ones, being those fadE16, fprA, Rv3230c, acs. EsxA, was downregulated, which is a major virulence factor related to regulation to the immune response of the host, which can be linked to the loss of virulence of the mutant M.tb. cells. We also found some genes related to cell-wall processes (Rv1978, rv3137), which again hints at a change in the characteristics of the cell wall. This idea is supported by looking at some of the upregulated genes in the KO condition. Rv3717, a protein related to cell wall degradation, is upregulated in the mutant, which can be linked to the higher permeability of the VirKO strain. Several other membrane related proteins, such as rv0669c, a protein involved in the synthesis of sphingolipids, a type of membrane lipids, and rmlA, which is involved in the biosynthesis of extracellular polysaccharides, are upregulated in the mutant too.

Two translation-related proteins, yfiA and rplQ, were found upregulated in the mutant. In the case of rplQ, it was also found upregulated in the transcriptomics results (Table 2). Some other proteins were related to amino-acid biosynthesis (thiC, hsaF, hisG) and regulation of transcription (pcnA, rho, rv3716c). We also found some upregulated proteins related to detoxification (mca) and adhesion to host (psts1).

	logFC in transcriptomic data	**FDR in transcriptomic data**	**logFC in proteomic data**	**FDR in proteomic data**
rplQ	0,107651	0,022407	3,551330987	0,09143

Table 2. logFC and FDR values of gene rplQ, for the transcriptomics and proteomics results in WC

For the case of the set of differentially expressed proteins found in the WT vs KO contrast in the extracellular vesicle location, the larger size allowed for a separate study of the upregulated and downregulated terms in the KO condition. Thanks to that, we could produce the following ClueGO networks:

Downregulated terms in KO EVs

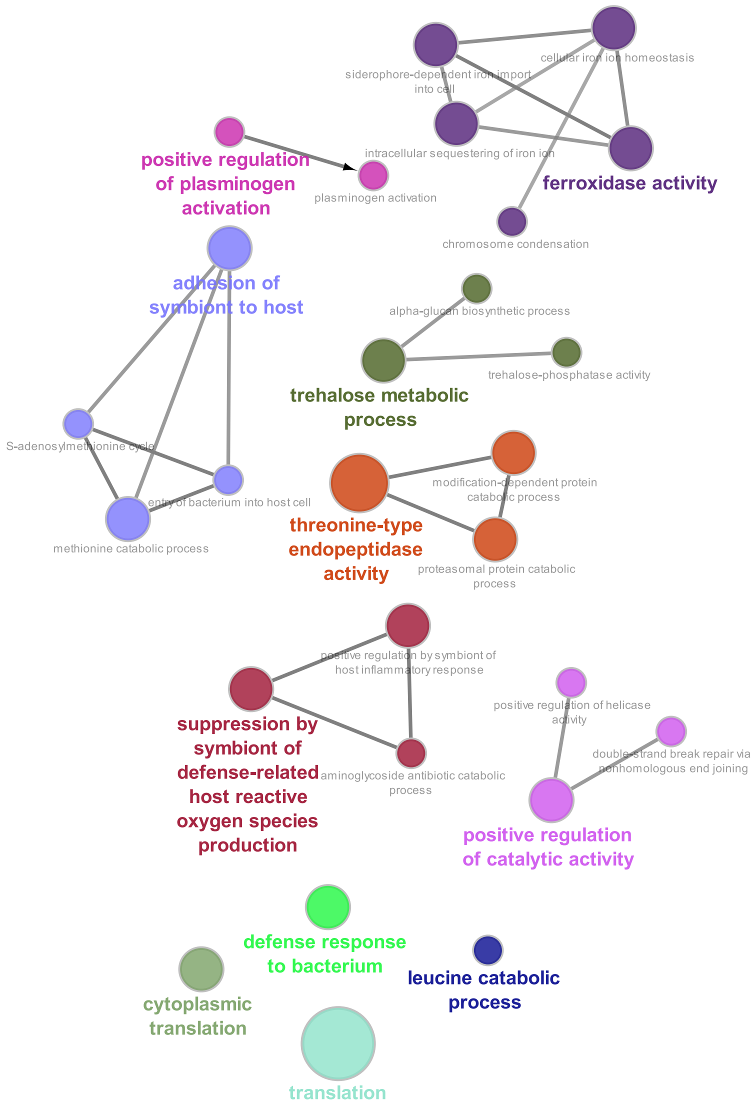

siderophore-dependent iron import into cell
cellular iron ion homeostasis
intracellular sequestering of iron ion
ferroxidase activity
chromosome condensation
positive regulation of plasminogen activation
plasminogen activation
adhesion of symbiont to host
alpha-glucan biosynthetic process
trehalose-phosphatase activity
trehalose metabolic process
S-adenosylmethionine cycle
entry of bacterium into host cell
methionine catabolic process
modification-dependent protein catabolic process
threonine-type endopeptidase activity
proteasomal protein catabolic process
positive regulation by symbiont of host inflammatory response
suppression by symbiont of defense-related host reactive oxygen species production
aminoglycoside antibiotic catabolic process
positive regulation of helicase activity
double-strand break repair via nonhomologous end joining
positive regulation of catalytic activity
defense response to bacterium
leucine catabolic process
cytoplasmic translation
translation

Figure 20. Enrichment network of the differentially expressed proteins of the extracellular vesicles upregulated in the WT condition.

Among the terms associated to the downregulated proteins in extracellular vesicles, we found several enrichments in processes which agree with our previous knowledge of the associated functions of extracellular vesicles. These terms include the ones related to defense response to bacterium, suppression by symbiont of defense-related host reactive oxygen species production, adhesion of symbiont to host, and ferroxidase activity. All these proteins are less expressed in EVs from VirR KO, indicating that the EVs of the mutant strain are individually less efficient in the deployment of the mentioned biological processes. We also found a large enrichment in translation related proteins, suggesting that the production of these proteins may be regulated in situ, within ribosome-containing vesicles. These results are coherent with the thought of EVs being involved in carrying virulence factors and molecules involved in the cross-talk between M.tb. and host cells (4).

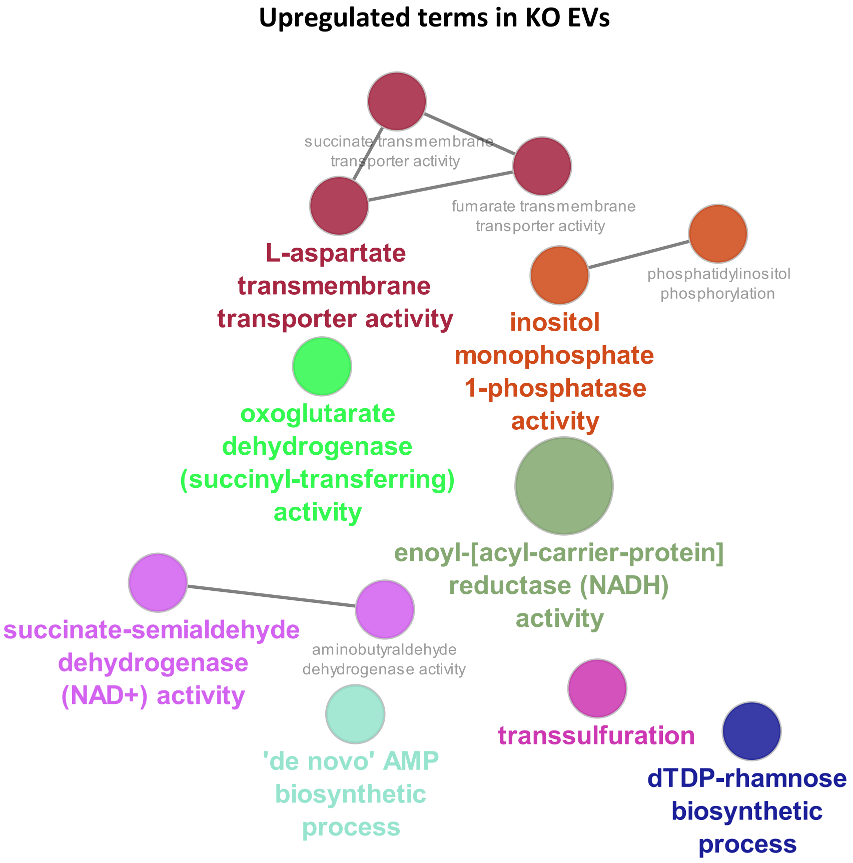

Upregulated terms in KO EVs
succinate transmembrane transporter activity
fumarate transmembrane transporter activity
L-aspartate transmembrane transporter activity
phosphatidylinositol phosphorylation
inositol monophosphate 1-phosphatase activity
oxoglutarate dehydrogenase (succinyl-transferring) activity
enoyl-[acyl-carrier-protein] reductase (NADH) activity
succinate-semialdehyde dehydrogenase (NAD+) activity
aminobutyraldehyde dehydrogenase activity
transsulfuration
'de novo' AMP biosynthetic process
dTDP-rhamnose biosynthetic process

Figure 21. Enrichment network of the differentially expressed proteins of the extracellular vesicles upregulated in the WT condition.

In the case of the upregulated terms in the KO, we find several terms related to metabolism of nucleotides, such as biosynthesis of AMP and phosphate-related activities on inositol. Other terms were related to reductase and dehydrogenase activities, and several transmembrane transporting activities. This indicates that protein cargo related to cross-talk has been widely lost, which means that the absence of VirR is affecting the function of the vesicles. M.tb. cells with a non-functional copy of VirR have been observed by our collaborators to increase the amount of secreted extracellular vesicles. Knowing that extracellular vesicles suffer a great loss of function when looking at their cargo in the KO condition, the increase of extracellular vesicles production may be related to an attempt to compensate for the low effectivity of the EVs by increasing their production.

7.7. Integration analysis

For the integration analysis, we constructed two different enrichment tables, one for each direction of differential expression. These tables, as well as the corresponding Fisher tests and a summary plot of the odds-ratio of each test, are presented here:

Upregulated terms	Found in proteomic analysis	Not found in proteomic analysis
Found in transcriptomic analysis	6	154
Not found in transcriptomic analysis	15	748

P-value: 0.2363796 Odds Ratio: 1.941266

Table 3. Contingency table of upregulated proteins found in either proteomics, transcriptomics, both, and neither set.

Downregulated terms	Found in proteomic analysis	Not found in proteomic analysis
Found in transcriptomic analysis	3	256
Not found in transcriptomic analysis	4	660

P-value: 0.1942765 Odds Ratio: 2.016999

Table 4. Contingency table of downregulated proteins found in either proteomics, transcriptomics, both, and neither set.

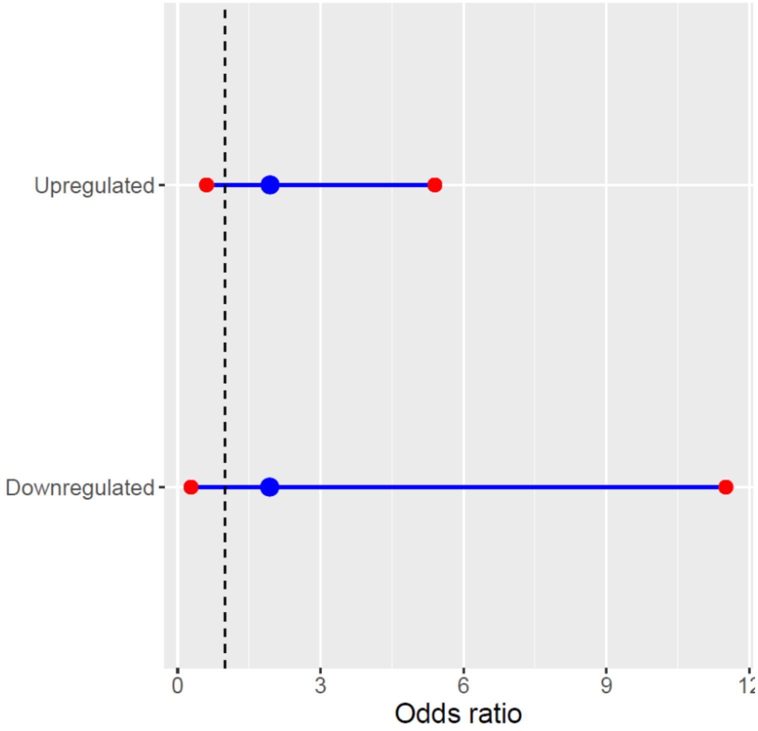


Figure 22. Odds ratio from the Fisher Exact Test performed on each group, and its corresponding confidence interval. The position of the dashed line indicates Odds ratio = 1.

The odds ratio relates the size of the observed intersection and its expected-by-chance size. An odds ratio greater than 1 indicates that the intersection is greater than expected at random, which means there is enrichment between the sets. On the other hand, an odds ratio lower than 1 indicates a lower than expected at random intersection, meaning that there is depletion between the sets.

Neither of the tests embraces the alternative hypothesis ($OR \neq 1$), which implies that there is no significant concordance between the differentially expressed genes and proteins. Further comparisons were not possible due to the fact that the provided transcriptomics and proteomics data were not obtained from the same cultures, although being exposed to the same conditions. The high number of translation-related genes found differentially expressed in transcriptomics data suggest that post-translational regulation has a clearly relevant role in the regulation of homeostasis after removing VirR, and that such extra regulatory layer essentially decouples the transcriptomic from the proteomic profiles.

8. Conclusions

The analyses conducted in this Final Master Project delineate a general portrait of the systemic regulatory role of the VirR protein. VirR is involved in the regulation of many processes in M.tb., and its absence greatly affects the transcriptomic landscape of the cells, suggesting an important role in regulatory pathways related to permeability and detoxification, metabolism and translation. Moreover, several of these processes are related to cell wall components, which supports the idea of VirR having an important role in cell wall maintaining processes. The transcriptomic analyses here presented indicate that VirR also has a determinant role in modulating the virulence of M.tb. cells, as its absence leads to downregulation of virulence related genes, which in turn leads to a decreased virulent capacity in the VirR-KO strain.

VirR has a very important role in the determination of the protein cargo of extracellular vesicles, too. Vesicles secreted by cells with a non-functional copy of VirR show largely modified proteomic profiles, even though the low sample size of the study only allows us to describe N=84 DEPs between KO and WT strains in their EVs. The vesicles from the VirR-KO strain present an underrepresentation of iron metabolism and interaction with host related genes when compared to vesicles secreted by WT cells, pointing toward a primary involvement of EVs in mediating virulence in M. tb.. In this context, the observed increase in vesicle production in VirR mutant cells, which show less functional proteomic profiles in their EVs, could be a way of the mutant cells to try to compensate for the loss of function of their vesicles.

Finally, the regulation of these processes has a great post-transcriptional component, as ascertained from the upregulation of translation and RNA modification related terms in the mutant strain and the lack of significant coincidence between the sets of differentially expressed genes and proteins from the whole cell samples. In order to fully interpret these results, further experiments are required, including transcriptomic analyses on EVs alone, analyses done on larger sample sizes, enabling simultaneous interrogation of proteomic and transcriptomic data from the same cells (or at least, the same cultures) and (or) analyses -either transcriptomic or proteomic- of vesicle depleted whole cell extracts.

Some of these new experiments are already planned, and we expect to use them to shed light on the biology of cell wall biogenesis and vesiculation in M.tb.

9. Bibliography

1. who.int. Tuberculosis [Internet]. 2022 [cited 2023 Jan 30]. Available from: https://www.who.int/news-room/fact-sheets/detail/tuberculosis

2. Russell DG. Mycobacterium tuberculosis: here today, and here tomorrow. Nature Reviews Molecular Cell Biology 2001 2:8 [Internet]. 2001 Aug [cited 2022 Nov 21];2(8):569–78. Available from: https://www.nature.com/articles/35085034

3. Roy A, Eisenhut M, Harris RJ, Rodrigues LC, Sridhar S, Habermann S, et al. Effect of BCG vaccination against Mycobacterium tuberculosis infection in children: systematic review and meta-analysis. BMJ. 2014 Aug 5;349(aug04 5):g4643–g4643.

4. Rodriguez GM, Prados-Rosales R. Functions and importance of mycobacterial extracellular vesicles. Appl Microbiol Biotechnol. 2016 May 29;100(9):3887–92.

5. Magombedze G, Dowdy D, Mulder N. Latent Tuberculosis: Models, Computational Efforts and the Pathogen's Regulatory Mechanisms during Dormancy. Front Bioeng Biotechnol. 2013;1.

6. Ehrt S, Schnappinger D. Mycobacterium tuberculosis virulence: lipids inside and out. Nat Med. 2007 Mar;13(3):284–5.

7. Delogu G, Sali M, Fadda G. The Biology of Mycobacterium Tuberculosis Infection. Mediterr J Hematol Infect Dis [Internet]. 2013 [cited 2022 Nov 21];5(1):2013070. Available from: /pmc/articles/PMC3867229/

8. Alderwick LJ, Harrison J, Lloyd GS, Birch HL. The Mycobacterial Cell Wall—Peptidoglycan and Arabinogalactan. Cold Spring Harb Perspect Med. 2015 Aug;5(8):a021113.

9. ABRAHAMS KA, BESRA GS. Mycobacterial cell wall biosynthesis: a multifaceted antibiotic target. Parasitology. 2018 Feb 15;145(2):116–33.

10. Brown L, Wolf JM, Prados-Rosales R, Casadevall A. Through the wall: Extracellular vesicles in Gram-positive bacteria, mycobacteria and fungi. Vol. 13, Nature Reviews Microbiology. Nature Publishing Group; 2015. p. 620–30.

11. Olive AJ, Sassetti CM. Metabolic crosstalk between host and pathogen: sensing, adapting and competing. Nature Reviews Microbiology 2016 14:4 [Internet]. 2016 Mar 7 [cited 2022 Nov 21];14(4):221–34. Available from: https://www.nature.com/articles/nrmicro.2016.12

12. Zou C, Zhang Y, Liu H, Wu Y, Zhou X. Extracellular Vesicles: Recent Insights Into the Interaction Between Host and Pathogenic Bacteria. Front Immunol. 2022 May 25;13.

13. Mohammadzadeh R, Ghazvini K, Farsiani H, Soleimanpour S. Mycobacterium tuberculosis extracellular vesicles: exploitation for vaccine technology and diagnostic methods. Vol. 47, Critical Reviews in Microbiology. Taylor and Francis Ltd.; 2021. p. 13–33.

14. Mehaffy C, Kruh-Garcia NA, Graham B, Jarlsberg LG, Willyerd CE, Borisov A, et al. Identification of Mycobacterium tuberculosis Peptides in Serum Extracellular Vesicles from Persons with Latent Tuberculosis Infection. J Clin Microbiol. 2020 May 26;58(6).

15. Prados-Rosales R, Carreño LJ, Batista-Gonzalez A, Baena A, Venkataswamy MM, Xu J, et al. Mycobacterial Membrane Vesicles Administered Systemically in Mice Induce a Protective Immune Response to Surface Compartments of Mycobacterium tuberculosis. mBio. 2014 Oct 31;5(5).

16. Ballister ER, Samanovic MI, Darwin KH. Mycobacterium tuberculosis Rv2700 Contributes to Cell Envelope Integrity and Virulence. J Bacteriol. 2019 Oct;201(19).

17. Rodriguez GM, Prados-Rosales R. Functions and importance of mycobacterial extracellular vesicles. Appl Microbiol Biotechnol. 2016 May 29;100(9):3887–92.

18. Rath P, Huang C, Wang T, Wang T, Li H, Prados-Rosales R, et al. Genetic regulation of vesiculogenesis and immunomodulation in *Mycobacterium tuberculosis*. Proceedings of the National Academy of Sciences. 2013 Dec 3;110(49).

19. Rath P, Huang C, Wang T, Wang T, Li H, Prados-Rosales R, et al. Genetic regulation of vesiculogenesis and immunomodulation in *Mycobacterium tuberculosis*. Proceedings of the National Academy of Sciences. 2013 Dec 3;110(49).

20. Zhu W, Smith JW, Huang CM. Mass Spectrometry-Based Label-Free Quantitative Proteomics. J Biomed Biotechnol [Internet]. 2010 [cited 2022 Nov 21];2010. Available from: /pmc/articles/PMC2775274/

21. Välikangas T, Suomi T, Elo LL. A systematic evaluation of normalization methods in quantitative label-free proteomics. Brief Bioinform. 2018 Jan 1;19(1):1–11.

22. Lazar C, Gatto L, Ferro M, Bruley C, Burger T. Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies. J Proteome Res. 2016 Apr 1;15(4):1116–25.

23. Noble WS. How does multiple testing correction work? Nature Biotechnology 2009 27:12 [Internet]. 2009 Dec [cited 2022 Nov 21];27(12):1135–7. Available from: https://www.nature.com/articles/nbt1209-1135

24. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, et al. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics. 2009 Apr 15;25(8):1091–3.

25. Universidad Autónoma de Madrid. RAFAEL CARLOS PRADOS ROSALES [Internet]. [cited 2023 Feb 1]. Available from: https://portalcientifico.uam.es/es/ipublic/researcher/315002

26. Rens C, Chao JD, Sexton DL, Tocheva EI, Av-Gay Y. Roles for phthiocerol dimycocerosate lipids in Mycobacterium tuberculosis pathogenesis. Microbiology (N Y). 2021 Mar 1;167(3).

27. Mudunuri U, Che A, Yi M, Stephens RM. bioDBnet: the biological database network. Bioinformatics. 2009 Feb 15;25(4):555–6.

28. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014 Dec 5;15(12):550.

29. Zhang X, Smits AH, van Tilburg GB, Ovaa H, Huber W, Vermeulen M. Proteome-wide identification of ubiquitin interactions using UbIA-MS. Nat Protoc. 2018 Mar 15;13(3):530–50.

30. Goeminne LJE, Sticker A, Martens L, Gevaert K, Clement L. MSqRob Takes the Missing Hurdle: Uniting Intensity- and Count-Based Proteomics. Anal Chem. 2020 May 5;92(9):6278–87.

31. Sticker A, Goeminne L, Martens L, Clement L. Robust Summarization and Inference in Proteome-wide Label-free Quantification. Molecular & Cellular Proteomics. 2020 Jul;19(7):1209–19.

32. Goeminne LudgerJE, Gevaert K, Clement L. Peptide-level Robust Ridge Regression Improves Estimation, Sensitivity, and Specificity in Data-dependent Quantitative Label-free Shotgun Proteomics. Molecular & Cellular Proteomics. 2016 Feb;15(2):657–68.

33. Raffelsberger W. wrProteo: Proteomics Data Analysis Functions. R package version 1.7.0.1, <https://CRAN.R-project.org/package=wrProteo>. 2022.

34. Huber W, von Heydebreck A, Ultmann HS¨, Poustka A, Vingron M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression [Internet]. Vol. 18, BIOINFORMATICS. 2002. Available from: http://www.dkfz.de/abt0840/whuber

35. Oba S, Sato MA, Takemasa I, Monden M, Matsubara KI, Ishii S. A Bayesian missing value estimation method for gene expression profile data. Bioinformatics. 2003 Nov 1;19(16):2088–96.

36. Hamid Z, Zimmerman KD, Guillen-Ahlers H, Li C, Nathanielsz P, Cox LA, et al. Assessment of label-free quantification and missing value imputation for proteomics in non-human primates. BMC Genomics. 2022 Dec 1;23(1).

37. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015 Apr 20;43(7):e47–e47.

38. Strimmer K. fdrtool: a versatile R package for estimating local and tail area-based false discovery rates. Bioinformatics. 2008 Jun 15;24(12):1461–2.

39. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, et al. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics. 2009 Apr 15;25(8):1091–3.

40. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res. 2003 Nov;13(11):2498–504.

41. Carbon S, Douglass E, Good BM, Unni DR, Harris NL, Mungall CJ, et al. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res. 2021 Jan 8;49(D1):D325–34.

42. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000 May;25(1):25–9.

43. Quigley J, Hughitt VK, Velikovsky CA, Mariuzza RA, El-Sayed NM, Briken V. The Cell Wall Lipid PDIM Contributes to Phagosomal Escape and Host Cell Exit of *Mycobacterium tuberculosis*. mBio. 2017 May 3;8(2).

44. Upton AM, McKinney JD. Role of the methylcitrate cycle in propionate metabolism and detoxification in Mycobacterium smegmatis. Microbiology (N Y). 2007 Dec 1;153(12):3973–82.

45. Bateman A, Martin MJ, Orchard S, Magrane M, Ahmad S, Alpi E, et al. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 2023 Jan 6;51(D1):D523–31.