Master Thesis

Identification of novel *Mycobacterium tuberculosis* proteins expressed *in vivo* in human pulmonary tuberculosis lesions: meta-proteomics of the autopsy material from pre-antibiotic era, 1931 - 1947.

Beatrice Normann

Centre for International Health, Department of Global Public Health and Primary Care
The Proteomics Unit at the University of Bergen (PROBE), Department of Biomedicine
Faculty of Medicine
University of Bergen, Norway
2022

Master Thesis

Identification of novel *Mycobacterium tuberculosis* proteins expressed *in vivo* in human pulmonary tuberculosis lesions: meta-proteomics of the autopsy material from pre-antibiotic era, 1931 - 1947.

Beatrice Normann

This thesis is submitted in partial fulfilment of the requirements for the degree of Master of Philosophy in Global Health at the University of Bergen.

Centre for International Health, Department of Global Public Health and Primary Care
The Proteomics Unit at the University of Bergen (PROBE), Department of Biomedicine
Faculty of Medicine
University of Bergen, Norway
2022

Abstract

Tuberculosis (TB) is an infectious disease caused by the bacteria, *Mycobacterium tuberculosis*. In recent years, research on proteins and peptides have been increasing as their potential for preventative and therapeutic treatment is becoming more attractive in tackling diseases. We propose to use a meta-proteomics approach to detect *Mycobacterium tuberculosis* proteins present in vivo in the lesions of human pulmonary TB. The early stages of disease have very low burden of *Mycobacterium tuberculosis* and these bacterial proteins are not detectable by using routine detection methods such as acid-fast staining or immunohistochemistry. The in vivo expressed proteins in the early stages of disease are of interest to develop novel vaccines. We aimed to establish the methodology to detect *Mycobacterium tuberculosis* proteins in the TB lesions at different stages of disease by proteomic analysis of the laser-captured micro-dissected lesions despite paucity of microscopically detectable bacilli and antigens.

Autopsy material from five lungs, a range of death between 1935 – 1947, with a total of eleven samples were used. Six early lesions and five necrotic lesions were identified, micro-dissected and processed for the proteomic analysis by mass spectrometry. A total of 3531 *homo sapiens* proteins and 110 bacterial proteins identified. Of these bacterial proteins, 5 were mycobacterial proteins; 1 *Mycobacterium avium* and 4 *Mycobacterium tuberculosis*.

The protein, RV2971, was significantly differentially expressed in the early lesion compared to the necrotic lesions and mihF was significantly differentially expressed in the necrotic lesions as compared to the early lesions. The proteins groEL2, hup and cycA were not significantly differentially expressed, however had a slightly higher interaction in the necrotic lesions than the early lesions.

In conclusion, this pilot study has established a proof of principle that it is feasible to detect and identify very low quantities of mycobacterial proteins present in the abundant host proteins, which are otherwise not detectable by routine methods. This methodology shows promise in strengthening theories and bringing forth new knowledge that could be used to build, improve and advance medicine for TB across the globe.

Table of Contents

Acronyms and Abbreviations

- TB: Tuberculosis
- MTB: *Mycobacterium tuberculosis*
- WHO: World Health Organisation
- HIV: Human immunodeficiency virus
- TST: Tuberculin skin test
- IGRA: Tuberculosis blood test
- PCR: Polymerase chain reaction
- FFPE: Formalin-fixed paraffin embedded
- BCG: Bacille Calmette Guérin
- MTB-TB: Multidrug resistant tuberculosis
- MS/MS: Tandem mass spectrometry
- m/z: Mass-to-charge ratio
- H2O: Water
- TCEP: Tris(2-carbocyethyl) phosphine
- CAA: Chloroacetamide
- BCA: Bicinchoninic acid
- AmBic: Ammonium bicarbonate
- CaCl2: Calcium chloride
- HCL: Hydrochloric acid
- NaCl: Sodium chloride
- TFA: Trifluoracetic
- ACN: acetonitrile
- HE: Haematoxylin eosin
- emPAI: Exponentially modified protein abundance index
- PSM: Peptide-spectrum match
-

Acknowledgements

I feel incredibly lucky to have such an amazing, kind and supportive team on my side throughout this thesis. Without any of them, I do not know what I would have done.

Firstly, I would like to thank professor Tehmina Mustafa from the Faculty of Medicine at the University of Bergen for being the main supervisor for this project. She gave me this incredible opportunity and trusted me to work with such valuable material. She encouraged my potential, inspired my passion and kept me positive through it all.

PhD candidate, Syeda Mariam Riaz, from the Faculty of Medicine at the University of Bergen was my co-supervisor for this project. She was incredibly patient in her teachings, had a beautiful baby girl and still managed to be there for me when I needed her.

Head engineer, Even Birkland, from the Department of Biomedicine at the University of Bergen, was my second co-supervising for this project. He guided me through proteomics, reassured me when I felt lost and created an environment that felt both safe and fun.

Another thanks to senior technician, Bendik Nordanger, from Klinisk Institutt 1 at Haukeland University Hospital for preparing the samples onto PEN-membrane glass slides and senior engineer, Kenneth Finne, from Klinisk Institutt 1 at Haukeland University Hospital for teaching me the operation of the laser microdissection. I would also like to especially thank the PROBE team at the University of Bergen, Olav Mjaavantten, Paula Carrillo Rodriguez, Harald Barsnes and Frode Selheim, for their assistance, advice and entertainment during my time at PROBE.

Finally, I would like to thank my sisters, Victoria Normann and Jessie Fu, my mother and my forever friends; Marius Antonsen, Misha Moghaddam, Amimah Gilani, Jennie Allred, Alaïs Peyrat, Ingrid Willumsen, Ahmed Zafeer, Gen Go and Jacqueline Lai for their endless love, humour and support. Lastly, Guillaume Brignoli, I love you always, I miss you with my whole soul and you will forever be my iglu brother.

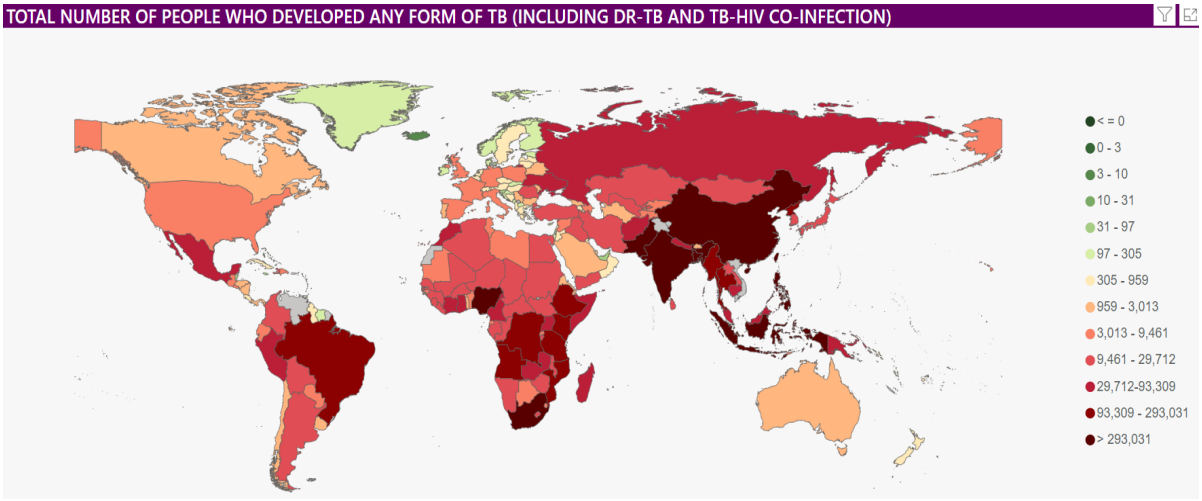
1. Introduction

1.1 Tuberculosis

1.1.1 Global impact

Tuberculosis (TB) is an infectious disease caused by the bacteria, *Mycobacterium tuberculosis* (MTB), with between 70-92% of patients having pulmonary involvement. Although technology and medicine have advanced greatly this last century, especially with the discovery of effective and affordable chemotherapy more than fifty years ago, TB continues to affect the population with high morbidity and mortality and remains a public health priority. The World Health Organisation (WHO) has reported 10 million new cases and 1.4 million deaths worldwide due to TB in 2020, with around a quarter of the world's population infected with MTB (Figure 1 and 2). Due to the emergence of drug-resistant strains and the increasing association with human immunodeficiency virus (HIV), TB is one of the leading causes of morbidity and mortality and is a high burden on countries in South-East Asia and Africa and is ranked the second leading cause of death by a single infectious agent after COVID-19.[1-4]

a.



b.

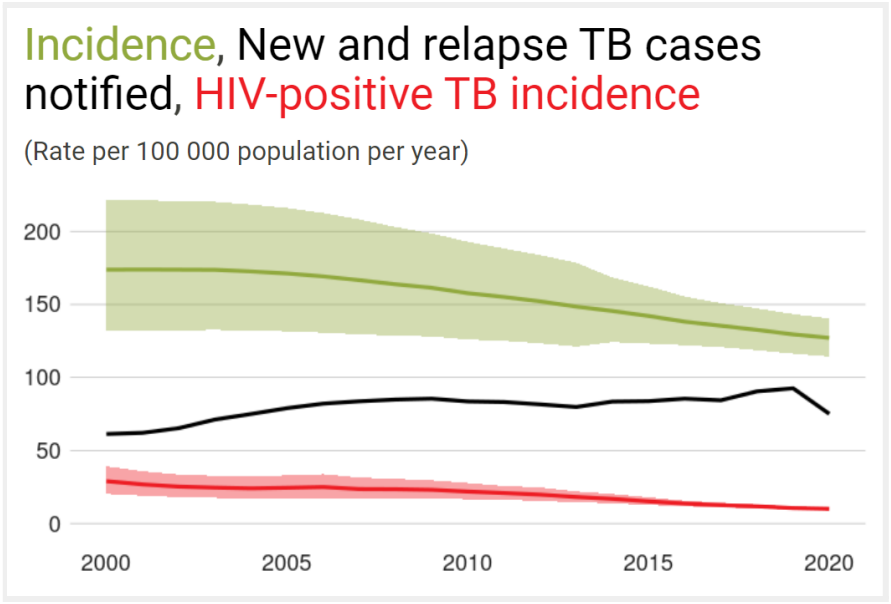
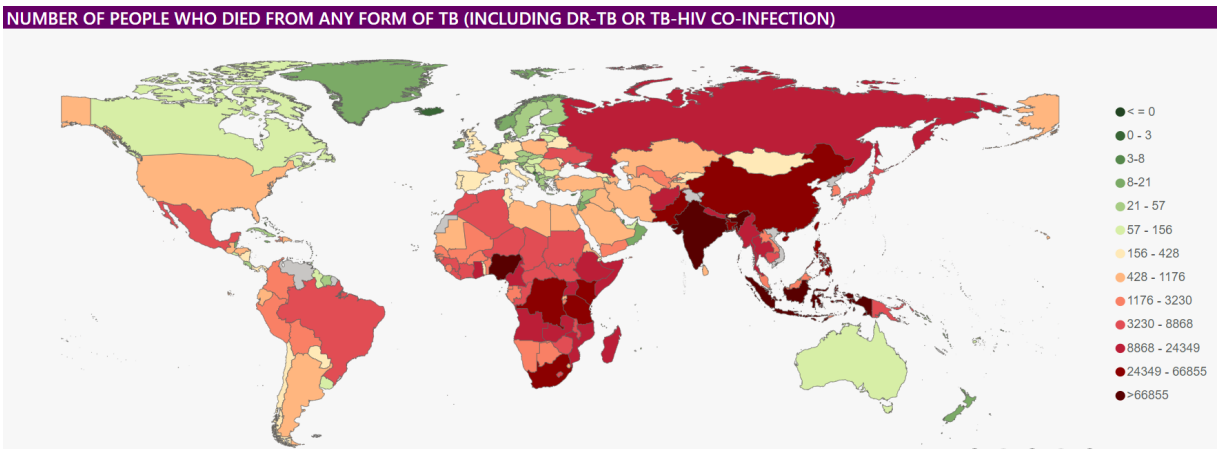


Figure 1: Incidence of TB. **a** | World map of the number of people who developed any form of TB, including drug-resistant TB or TB-HIV co-infection in 2020. Generated by Stop TB Partnership, hosted entity of the United Nations Office for Project Services[5]. **b** | Line graph of incidence rate (green line), new and relapse TB cases reported (black line) and HIV-positive TB incidence rate (red line) per 100,000 population per year globally from 2000 to 2020. Generated by the World Health Organisation[6].

a.



b.

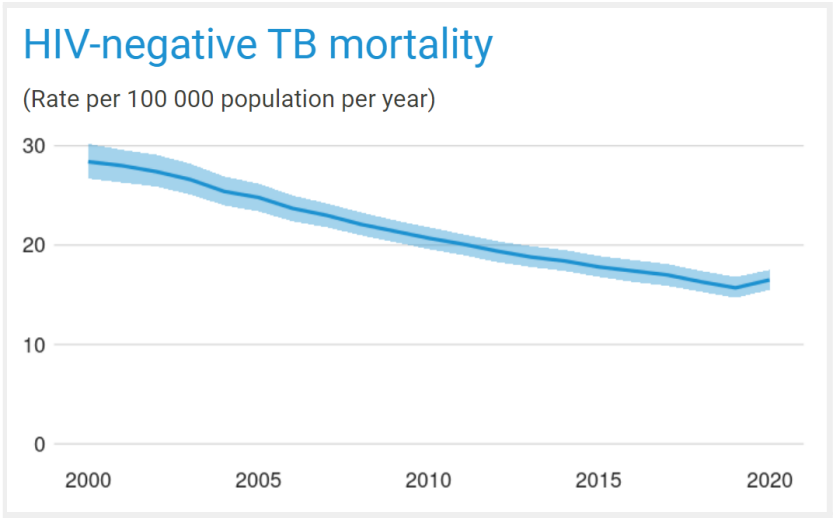


Figure 2: Mortality of TB. **a |** World map of the number of people who died from any form of TB, including drug-resistant TB or TB-HIV co-infection in 2020. Generated by Stop TB Partnership, hosted entity of the United Nations Office for Project Services.[7] **b |** Line graph of TB mortality rate per 100,000 population per year globally from 2000 to 2020. Generated by the World Health Organisation[8].

1.1.2 Sustainability Developmental Goals

The United Nations created seventeen Sustainability Developmental Goals. TB falls under the third Sustainability Development Goal; good health and well-being. Specifically, target 3.3 states that "by 2030, end the epidemics of AIDS, tuberculosis, malaria and neglected tropical diseases and combat hepatitis, water-borne diseases and other communicable diseases" and this project hopes to assist in fulfilling this target.[9]

1.1.3. Pathogenesis of primary, post-primary and latent TB

TB is divided into primary, post-primary and latent TB. Primary TB occurs when the individual is infected with MTB for the first time without prior immunity to the infection. About 85 - 95% of immunocompetent individuals are able to mount a protective immune response against MTB and restrict the infection leading to the formation of a granuloma. Granuloma is a characteristic of primary TB. They are small inflammations that often form in the lungs but can form anywhere in the body. The formation of granuloma indicates an effective immune response mediated by the

adaptive immunity, which is able to control the infection, but seldom eradicates the infection, resulting in dormant infection [10]. The progression to clinical disease occurs in very few, typically in immunosuppressed or extremes of ages. The primary TB manifests as disseminated disease especially in immunocompromised and young, with less cavitation and transmission and more mortality [11].

Latent infection occurs when the person harbours MTB in the body but does not present any signs or symptoms of the clinical disease. This can occur after a primary infection, or after primary disease, where the host adaptive immune response has been activated to control the pathogen and force it into a dormant state. Dormant MTB can remain alive within the human host for years. Although individuals with latent TB are unable to transmit the disease to others, they constitute a large reservoir of MTB that can cause an outbreak of the disease in the future.[12]

Post-primary TB only begins after primary MTB has established systemic immunity. It can occur when an individual becomes reinfected by the bacteria or TB is reactivated. Post-primary TB constitutes the main disease burden in adults. The disease is more commonly restricted to the lungs with cavity formation and individuals can transmit the disease, sustaining the disease within the community.

While MTB can infect many animals, they cannot transmit the infection. The continued survival of MTB, therefore, depends upon transmission among humans. This is best accomplished by producing a cavity in the lung for proliferation of massive numbers of MTB to be coughed into the environment over a period of decades while the host remains healthy enough to circulate in the community. Therefore, primary TB disease is not favourable for MTB and post-primary TB disease is favourable for MTB.

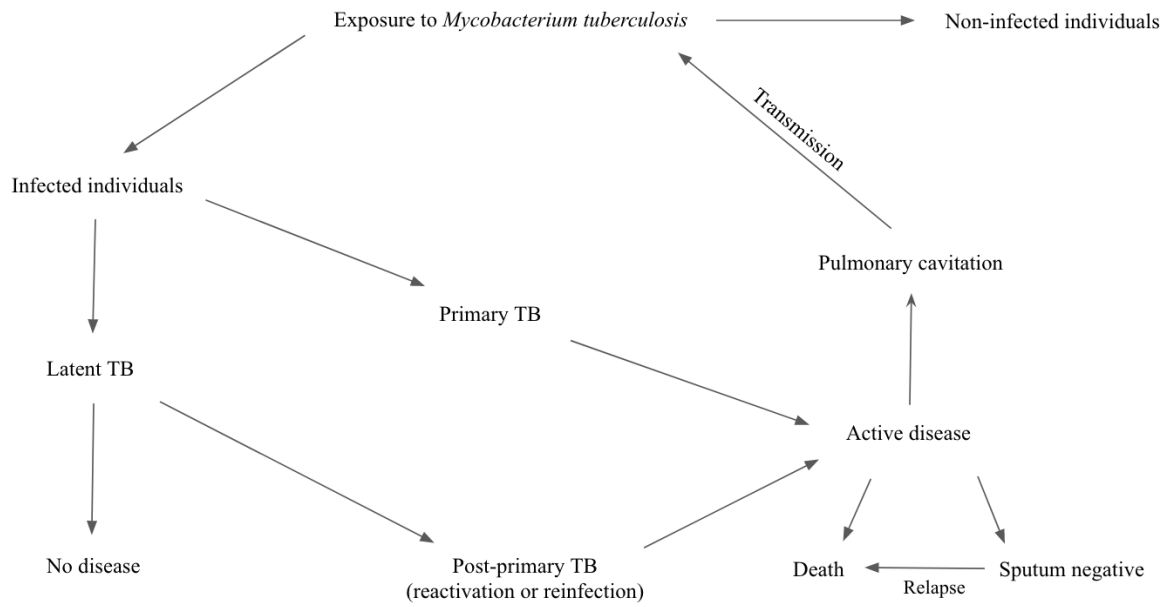


Figure 3: Natural course of TB without antibiotics.

Figure 3 is a diagram of the natural course of TB. It begins with an individual being exposed to MTB. About 85 - 95% of immunocompetent individuals are able to mount a protective immune response against MTB and restrict the infection leading to the formation of a granuloma. The progression to clinical disease occurs in very few, typically in immunosuppressed or extremes of ages. Primary TB manifests as disseminated disease especially in immunocompromised and young. Post-primary TB occurs as a result of reactivation or reinfection. Around 45% of patients with HIV-negative, active TB and all patients with HIV-positive, active TB will die in the absence of proper treatment. Otherwise, patients will receive a negative test on sputum smear and possible relapse or the patient will continue to form pulmonary cavities and transmit the disease. This process is a cycle and can take decades, which is why it is incredibly difficult to contain TB.[3]

1.1.4 Factors that increase the risk of tuberculosis

Anyone can develop TB, however, there are a number of factors that can increase the likelihood of contacting MTB and developing TB disease. As previously mentioned, TB disease can take

decades to develop, however in some cases, TB disease can develop weeks after infection due to a weak immune system. Infants and immunocompromised individuals are at risk to this.
Those with HIV and diabetes mellitus have a higher chance of developing the disease. An unhealthy diet, obesity or undernutrition, substance abuse and smoking can also cause the body to weaken, therefore developing the disease.[13]

1.1.5 Diagnosis

Diagnosis of TB is not straightforward and requires a combination of different tests in order to determine whether the disease is TB or not, if the disease is active or not and exactly what type of TB an individual has.

1.1.5.1 Signs and Symptoms

Common signs of TB disease are coughing, sputum, haemoptysis, breathlessness, weight loss, anorexia, fever, malaise, wasting and terminal cachexia. However, having these symptoms does not necessarily mean that it is TB and more testing is required to properly diagnose the disease, as these symptoms overlap with many other pulmonary infections and diseases.[14]

1.1.5.2 Microbiology

Microbiological confirmation of pulmonary TB is done by detection of bacteria from respiratory samples. The most common is sputum. Bronchoalveolar lavage and biopsy can also be obtained. MTB is detected by direct staining of smears by acid fast staining. Culture is the gold standard to detect if an individual has active TB. However, culture can take up to 8 weeks to produce results. Culture is considered better than sputum but in the majority of lower economically developed countries, sputum microscopy, known to have a lower specificity and sensitivity, is more accessible. This is an obvious concern as it can lead to more undetected active TB in patients, hence causing more difficulty in controlling the disease.[15-17]

1.1.5.3 Polymerase Chain Reaction (PCR) test

PCR tests detect TB by analysing the DNA or RNA from a sample and can amplify the sample using certain reagents. PCR tests are quick and have a 93% sensitivity, 84% specificity and the

ability to detect TB in formalin-fixed paraffin embedded (FFPE) blocks. However, due to its high cost, requirement to train personnel with the process and high maintenance of the equipment, it has not been placed in lower economically developed countries.[18-21]

1.1.5.4 Skin and blood test

TB skin test (TST) and TB blood test (IGRA) becomes positive if a person is infected with MTB, although it can only detect latent TB [22]. It should also be noted that TST can result in false-positives due to the *Bacille Calmette Guérin* (BCG) vaccine and IGRAs can result in false-negatives when the individual is immunocompromised, for example HIV, or TB is active [23]. WHO recommends TST as the main testing for latent TB in lower economically developed countries due to low cost and ease of use until better testing becomes available [24].

1.1.5.5 Chest radiography

Chest radiography or x-ray can detect abnormalities in the lungs. It is essential for early detection of TB, specifically pulmonary, as it can identify cavities, however, has a low specificity. Cavities are common in pulmonary TB but are also seen in pneumonia and lung cancer. Therefore, chest radiography alone is not conclusive of TB and can only show if the TB is or is not pulmonary TB.[15, 25]

1.1.6 Vaccine

The BCG vaccine was first administered to humans in 1921. It is the only vaccine that protects against primary TB. It has an 80% protection against common TB forms in children, such as TB meningitis and lasts up to 15 years. However, this does not give protection against post-primary TB. Today, the vaccine is only administered to infants from high-risk countries.[26]

1.1.7 Treatments

With the growing knowledge of TB, it is treatable with a 6 - 9 month treatment regimen that uses a combination of drugs with a risk of relapse of less than 5 - 8%. The drugs used are abbreviated as RIPE; Rifampin, Isoniazid, Pyrazinamide and Ethambutol. Treatment is complicated, with many side effects, and there is an emergence of multi-drug resistant strains which are difficult to

eradicate.[27, 28]

Rather than directly fighting the pathogen like antibiotics, host-directed therapy changes the environment the pathogen lives in [29]. Host-directed therapy has the potential to enhance anti-TB drug treatment, shorten the time of treatment and produce better treatment outcomes with fewer relapses by reducing inflammation, repairing or preventing tissue damage and regulating anti-MTB protective innate and adaptive immune system, using the host's own immune or bone marrow cells [30, 31]. It is known that the health of the host plays a huge contributing factor to the digression to TB disease and research on what exactly it is that causes this sudden change from latent to disease, could improve and help specify host-directed therapies.

1.1.8 Challenges

One reason why preventative strategies, particularly vaccines, are not good enough is due to the gap in knowledge. It is not yet established the correlation of protective immune response in post-primary TB nor has there been enough research into the difference of what is protective immune response and what causes tissue destruction and disease. To develop a better vaccine for post-primary TB, it must be known exactly what proteins and antigens are detected during disease reactivation and progress.

Currently, the biomarkers for treatment monitoring of pulmonary TB uses urine, blood or sputum samples and has a poor sensitivity [32, 33]. Therefore, it is essential that it is improved and updated. Though treating TB is possible, there are many challenges. Patient compliance is one of the major factors contributing to difficulty in treatment; as the treatment is complex and long, many patients quit therapy prematurely and this can then lead to the development of multidrug-resistant TB (MTB-TB). There is also the possibility of complications and toxicity with HIV and TB co-infections. Failed diagnosis is another problem, if the type of TB isn't properly diagnosed, the treatment will take longer and become more complicated. So, it is crucial that proper diagnosis is conducted. It has also been discussed in the scientific community that MTB is difficult to kill due to dormant cells which are less susceptible to antibiotics. Hence, why it takes long to treat as those dormant cells need to become active in order to kill them with

administered drugs.[34, 35]

Prevention and treatment of TB is a large problem in lower economically developed countries and as a result the incidence and death rate of TB is higher, see Figure 1 and 2 for reference. This is because of the difficulty of access to healthcare, high medical, testing and transport costs, low health knowledge, poor infrastructure and lack of proper and advanced medical equipment. Environmental, such as pollution and close proximity, lifestyle, such as physical inactivity and unhealthy diet, and geographical factors, such as living far from health facilities, also play a major role in the increased number of TB morbidity and mortality. Additionally, prisons and refugee camps are highly vulnerable to TB, as a consequence of crowded living conditions, poor nutrition, co-infection with other diseases and receiving less medical attention, with between 10 - 50% increased risk of developing TB compared to civilian sectors.[36-38]

Norway, a higher economically developed country, has recorded 3 cases per 100,000 people of TB in 2020 [39]. This shows that with proper and efficient preventative and therapeutic strategies in place, it is possible to control this disease.

1.2 Proteomics

Proteomics is a large-scale study of proteins in a biological system, such as a cell, tissue or organism. Using proteomics, researchers can quantify and identify proteins from a sample of interest. Proteomics from microbial communities is called meta-proteomics, which provides direct insight into the phenotypes of microorganisms on the molecular level [40, 41].
Two traditional methods used for proteomic analysis are the western blot and two-dimensional gel electrophoresis. Although western blot has a simple procedure, the reason western blot was not chosen for this project is due to the number of problems that arise from this method that requires troubleshooting [42]. Two-dimensional gel electrophoresis has poor consistency with results, faces difficulties when it comes to membrane proteins and has problems identifying low abundant proteins [43]. As samples that will be used for the project are valuable and inimitable, tandem mass spectrometry (MS/MS) was the method trusted and chosen to conduct the experiments.

1.2.1 The study of proteins and peptides

Proteins and peptides play a myriad of roles in the body [44]. The difference between the two is that peptides are shorter chains of two or more amino acids, whereas proteins are made up of multiple peptides, hence proteins are more diverse than peptides [45]. The interaction of different proteins with one another functions the body, so an increase or decrease in a certain protein could be catastrophic, leading to progression of diseases, including TB [46]. In recent years, research on proteins and peptides have been increasing as their potential for preventative and therapeutic treatment is becoming more attractive in tackling diseases [44, 47].

1.2.2 Mass spectrometry

A tool used in proteomics to identify, quantify and characterise molecules present in a sample is the mass spectrometer. Mass spectrometers are made up of three parts; an ion source, mass analyser and detector. When samples are put into the mass spectrometer, the ion source vaporises and ionises the samples, passing through the mass analyser which will measure the mass-to-charge ratio (m/z). Finally, it hits the detector, which collects the data and converts the number of ions at different m/z values into a spectra, ready for further analysis.[48-50]

1.2.3 Tandem mass tags (TMT)

Thermo Fisher Scientific designed Tandem Mass tag (TMT) labelling, which is a process of chemically labelling peptides, giving the ability to put all the samples together so that they can be run on the mass spectrometer together. This in turn helps in minimising biases that will be explored further in the discussion of this project. Each sample is given a “reporter ion” that has a different mass for the mass spectrometer to identify. MS2, which will be explained more in depth in the next section, quantifies TMT-tags by measuring the intensities of the reporter ions released from the labels, hence separating the samples.[51-54]

1.2.4 Tandem mass spectrometry (MS/MS)

Peptides are easier to sequence as they are smaller and database searching has more accuracy in matching unique amino acids to peptides [55]. An example of how MS and MS/MS will look like is displayed in Figure 4.

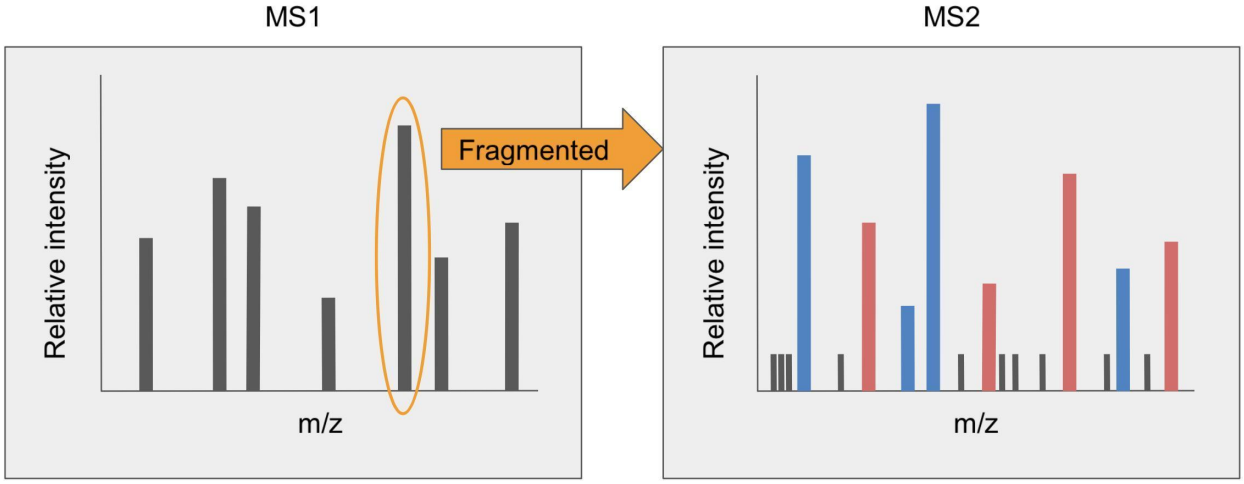


Figure 4: Example of data generated as MS and MS/MS. The y-axis is the relative intensity and the x-axis is the m/z value. In MS1, each line represents a peptide found in the sample and the one with the highest intensity is circled in orange. In MS2, each line represents fragments of the selected peptide from MS1. The lines in blue and red represent y- and b-ions, whereas those that are not coloured are considered interference.

MS1 displays the entire peptide and MS2 displays fragments of the selected peptide from MS1. Splitting the original peptide between various amino acids results in fragmented peptides; b- and y-ions. The b-ions extend from the amino terminus (N-terminus), whereas y-ions extend from the carboxyl terminus (C-terminus). y-ions are more common and more stable than b-ions. It is normal for samples that have TMT-labelling to have a higher interference in the beginning of MS2, as it will be detected by the mass spectrometer in order to separate the samples.[55-57]

1.3 Rationale of the study

In this master thesis, we propose to use a meta-proteomics approach to detect MTB proteins present in vivo in the lesions of human pulmonary TB. The early stages of disease have very low

burden of MTB and these bacterial proteins are not detectable by using routine detection methods such as acid-fast staining or immunohistochemistry. The in vivo expressed proteins in the early stages of disease are of interest to develop novel vaccines.

Currently, most clinical samples from TB patients are peripheral blood, lymph nodes, bronchoalveolar lavage cells, and lung biopsies. However, the information provided by these samples are limited. The responses and mechanisms examined in blood samples are systemic, not at the site of infection, and may represent secondary effects of the primary response in the lung tissue. The information from the alveolar lavages is limited to immune responses in the alveoli and biopsies samples are too small to incorporate the surrounding parenchyma. Furthermore, successful antibiotic treatment is known to significantly alter pathology. As MTB is killed by the antibiotic, the antigens that stimulate the host immune responses that generate the lung pathology are cleared. Thus, the most suitable material to understand the mechanisms of TB disease pathogenesis, especially the early lesions and the caseation process, would be the untreated human autopsy samples. [3, 4, 12, 29-31]

1.3.1 Norwegian archival material

The Gades Institute at the Department of Pathology, Haukeland University Hospital was established in 1912 and has since then been a referral centre for autopsy and biopsy diagnostics in Western Norway. Both biopsies and autopsy material have been stored systematically since 1930. This archival material is a precious, and one of its kind resource. Such material is not available in other countries, as most hospitals do not keep material for more than 10 – 20 years, and the records are not readily searchable. Norwegian archival material is stored in an organised manner, and combined with epidemiological surveys, patient's history, and detailed autopsy records provides us with a unique opportunity to investigate the pathogenesis of post-primary TB before the introduction of antibiotics. Previous studies by Lærum et al have shown that the material is optimally preserved, and modern research techniques can be used to study this archival material [58, 59]. Human T cell epitopes of MTB are evolutionarily, hyper conserved and, unlike some other viral and bacterial disease, MTB does not use antigenic diversity as a mechanism of immune evasion. Thus, the immune responses in the archived material will be relevant for the contemporary disease.

Previous studies have used the proteomics approach by mass spectrometry on the samples from FFPE material, which has been discussed to interfere with the identification of low level proteins when using proteomics [60]. Most of these studies have analysed the abundant host proteins, but there is a lack of studies on detection of extremely low levels of microbial proteins in the abundant host tissue proteins. In this project we aimed to establish the methodology to detect MTB proteins in the TB lesions at different stages of disease by proteomic analysis of the laser-captured micro-dissected lesions despite paucity of microscopically detectable bacilli and antigens.

1.3.2 Methodology

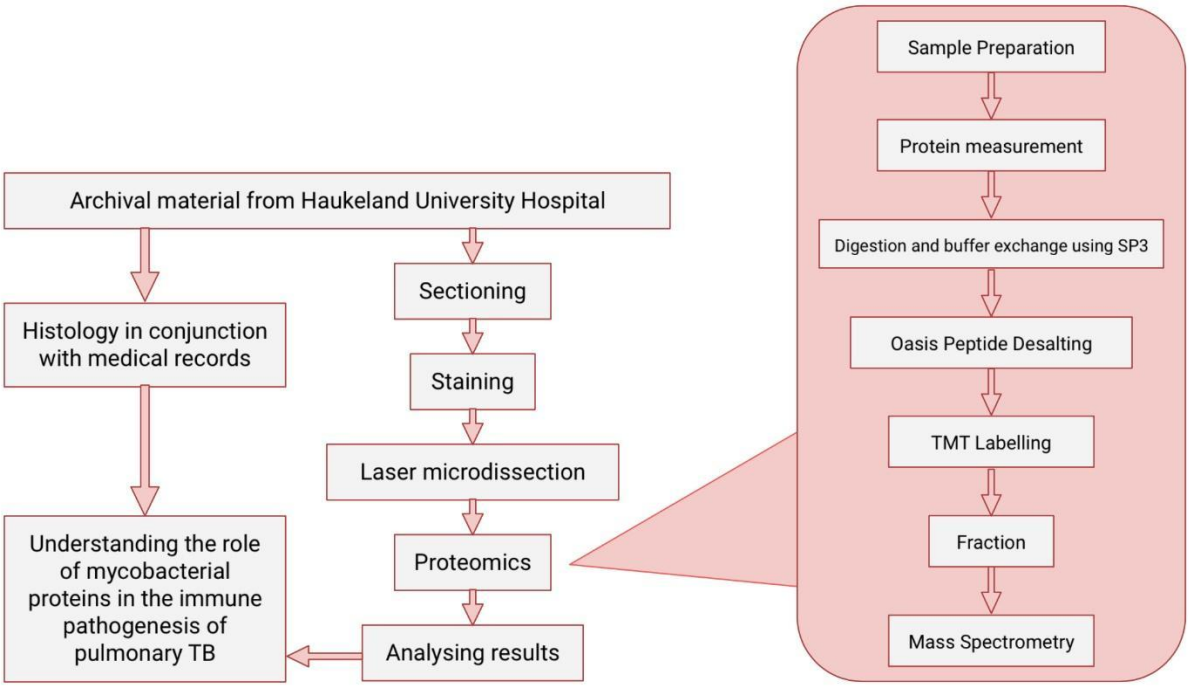


Figure 5: Methodology flow chart

A brief overview of the methodology used for this project is shown in Figure 5.

1.3.3 Questions, hypothesis and study objectives

There are five questions that have been explored in this project.

1. Was the methodology efficient and did it give valuable and new information?
2. Are FFPE blocks suitable for proteomics?
3. Which mycobacterial proteins were identified in the lesions?
4. Are there differences in proteins between the early and necrotic lesions?
5. Do the proteins interact and what is their role in the immune pathogenesis of TB?

The main objective of this project was to use proteomics to identify and quantify in vivo expressed mycobacterial proteins in the human TB disease at different stages of disease progression. The hypothesis of this study is to expect to find mycobacterial proteins and differences in expression within the lesions and similarities of the proteins between the same lesions among the different cases used. It is also hypothesised that the proteins found may interact with one another and will be associated with the progression of TB.

2. Methodology

Chemicals, equipment and software needed in order to conduct this experiment is listed in Table 1.

Table 1: Inventory list to conduct methodology.

Method	Inventory		
	Equipment	**Chemicals**	**Software**
Selecting cases	FFPE blocks of samples		
Sectioning of FFPE tissue blocks	Microtome Leica Biosystems, water bath, brushes, blades, PEN-membrane glass slides, hot cabinet		

Staining of tissues		Sigma-Aldrich Xylene, Sigma-Aldrich alcohol (100%, 96% and 80%), water (H2O), haematoxylin and liquid coverslips	
Laser microdissection of tissue sections	Leica LMD7 machine, Fisher Scientific standard thin-walled microcentrifuge tubes and their caps		
Proteomics: Sample preparation	Gilson Pipette, 1.5 mL Eppendorf tubes, Sonica sonicator, Eppendorf water bath, Fisher Scientific centrifuge	Sigma-Aldrich 4% SDS, Pierce QB Perbio 5mM tris(2-carboxyethyl) phosphine (TCEP), Sigma-Aldrich 10mM Chloroacetamide (CAA), VWR Chemicals 0.1 M Tris-HCL pH 8.5, VWR Chemicals H2O	
Proteomics: Protein measurement	Gilson pipette, Oasis 96 well plate with a flat bottom, Pechiney parafilm, Termaks hot cabinet, plate reader	Thermo Scientific Pierce BCA Assay Kit; Thermo Scientific Bicinchoninic acid (BCA) solution, Thermo Scientific Copper sulphate solution, VWR Chemicals H2O, Sigma-Aldrich Bovine	

		IgA/BSA, Thermo Scientific lysis buffer	
Proteomics: Digestion and buffer exchange using SP3	IKA vortex, Gilson pipette, Sansure Biotech Inc. magnetic rack, Eppendorf thermomixer, sonicator, 1.5 mL Eppendorf tubes, pH paper	Bead stock solution, Sigma-Aldrich 100% ethanol, Fluka Analytical ammonium bicarbonate (AmBic), VWR Chemicals H2O, Sigma-Aldrich calcium chloride (CaCl2), Sigma-Aldrich hydrochloric acid (HCL), Thermo Scientific trypsin, Fluka Chemicals sodium chloride (NaCl), Sigma-Aldrich 10% trifluoroacetic (TFA)	
Proteomics: Oasis peptide desalting	Oasis 96 well cartridge plate, waste plate, pipette, Fisher Scientific centrifuge, parafilm, -80 degree Celsius freezer, Labconco SpeedVac, shaker, Sonica sonicator	Sigma-Aldrich 80% acetonitrile (ACN), VWR 0.1% fluoroamphetamine (FA), Sigma-Aldrich 0.1% TFA, VWR 1% FA, Sigma-Aldrich 2% ACN	
Proteomics: TMT labelling	Gilson pipette, Eppendorf	Sigma-Aldrich 50% Hydroxylamine,	

	thermomixer, Thermo Scientific nanodrop, 1.5 mL Eppendorf tubes, Sonica sonicator, vortex, -80 degree Celsius freezer, Labconco SpeedVac	Sigma-Aldrich 50mM HEPES buffer, Thermo Scientific TMT reagents, Thermo Scientific TMT pro reagent, Sigma-Aldrich 0.1% TFA.	
Proteomics: Fractionation	2 mL Eppendorf tube, Fisher Scientific centrifuge, Gilson pipette, -80 degree Celsius freezer, Labconco SpeedVac, Eppendorf thermomix, Sonica sonicator, Fisher Scientific Nanodrop	Thermo Scientific Pierce High pH Reversed-Phase Peptide Fractionation Kit, Thermo scientific ACN, Sigma-Aldrich 0.1% triethylamine, Fluka 0.1% formic acid	
Proteomics: Mass spectrometer	Mass spectrometer	Blank sample, Helio sample	
Analysis			Uniprot, Thermo Proteome Discoverer version 5.2, Perseus version 1.6.15.0, String Database version 11.5, Cytoscape version

			3.9.1, Excel

2.1 Autopsy tissue material

FFPE blocks from 1931 – 1947 have been collected from the archives stored at the Department of Pathology, the Gades Institute, Norway as a part of another study. The inclusion criteria were all cases who had autopsy diagnosis or clinical diagnosis of any type of TB. The information for these cases were obtained from the autopsy journals containing all relevant medical history including patient characteristics, medical history, any procedures or diagnostic tests, diagnosis before death if any, autopsy diagnosis, and detailed gross description and brief microscopic description.

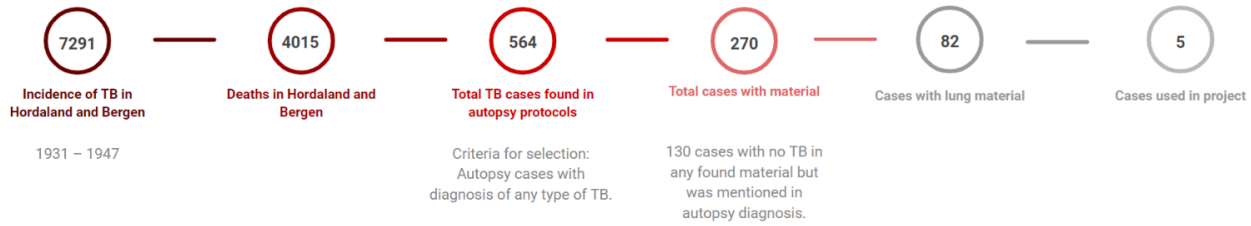


Figure 6: The incidence and mortality of TB in Bergen, Hordaland in the period between 1931-1947 and number of autopsy cases found in the autopsy reports.

Figure 6 shows the incidence and mortality of TB in Hordaland between 1931 – 1947, which were high in Norway during this time period [61]. From the 7291 incidences of TB, 4015 deaths were recorded and from these, a total of 564 TB cases were found in the autopsy reports. There were a total of 270 cases with available material, however, TB was not identified under a microscope from 130 of the cases with material but was mentioned in the autopsy reports. Of the cases with available material, 82 cases had lung material which were relevant to the project and 5 of these cases were used.

All cases selected for this project had pulmonary TB and were chosen not because of patient characteristics but due to visibility and ease of the ability to cut host cells and specific lesions.

2.2 Sectioning of FFPE tissue blocks

Tissues were sectioned by using a microtome, cut at 10 micrometres, treated with UV light for 10 minutes before being mounted onto PEN membrane glass slides. PEN membrane glass slides were used because it has better consistency in capturing and collecting cells and is also inexpensive compared to other types of slides that are used for laser microdissection. Studies have found that PEN membrane slides are more reliable because they do not rely as much on the dehydration of the tissue, compared to the glass slides [62].

2.3 Staining of tissues prior to laser microdissection

Staining assists in visibility under a microscope to prepare for laser microdissection. Since this project will compare two different type lesions, early and necrotic, being able to identify them under the microscope is crucial. Therefore, haematoxylin eosin (HE) was used to stain the slides. Haematoxylin stains nuclear components, which aids in identifying patterns, shape and structure of the cells in the samples.

To conduct staining, slides were first deparaffinised and gradually rehydrated with alcohol. The slides were dipped into xylene and 100% alcohol twice for five minutes each. Then in 96% alcohol and 80% alcohol once for 2 minutes. The slides were then put into water for 2 minutes. Next 2-3 drops of haematoxylin (with a dilution of 1:10) were applied for 5 seconds and the slides were then washed with warm water. The hydration using alcohol was then done in reverse. Before the xylene dries, liquid coverslip was sprayed onto the slides 1-2 times (10 cm away from the slide) and dried for 5 minutes.

2.4 Laser microdissection of haematoxylin-stained TB FFPE tissue sections

To cut sections of each case, laser microdissection was used. This machine enabled the user to cut specific sections of a slide under a microscopic level and these microdissected cells were then used for protein profiling [63, 64].

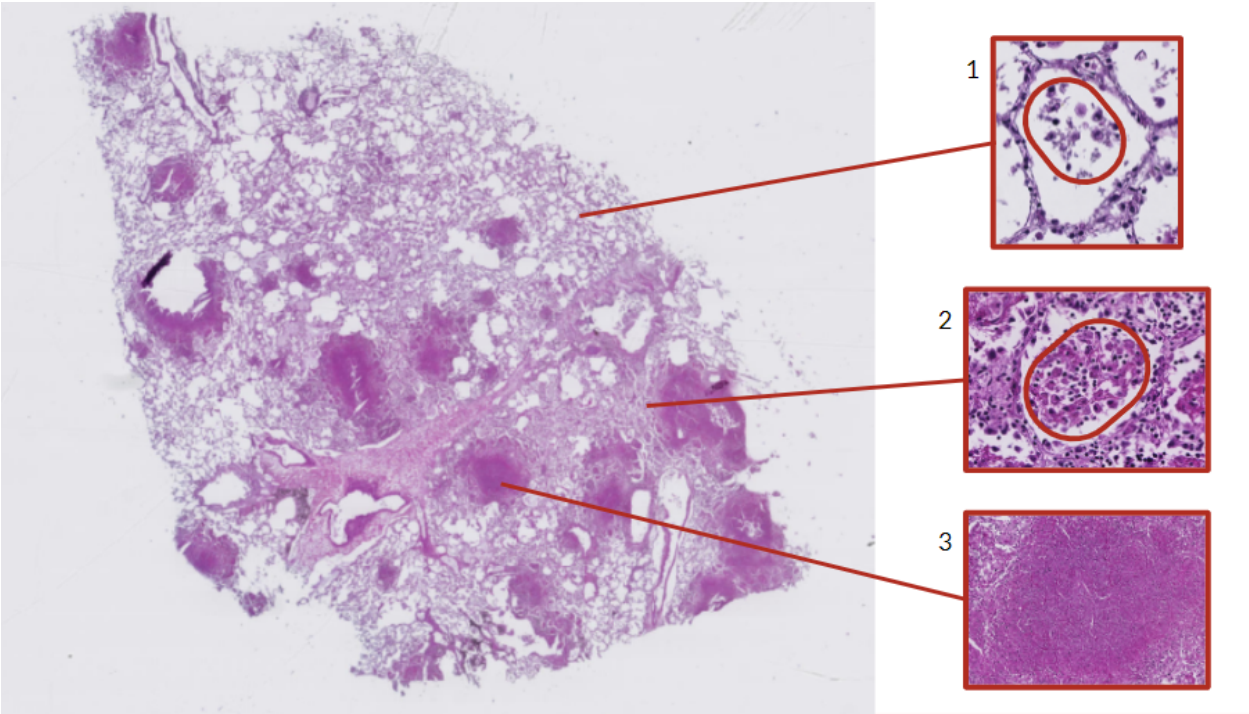


Figure 7: Scan of one of the cases used in this project.

Scans of a HE-stained slide of the cases were used for reference and a photo of one of the cases is shown in Figure 7. Next to it are three zoomed in photos of sections cut. Many sections were cut from each case. The sections were cut into two different groups: early lesions and necrotic lesions. 1 and 2 is an example of what were classified as early lesions and 3 was a necrotic lesion. Necrosis occurs when there is a lack of blood and oxygen in the cells, which leads to the cells becoming acidic and enzymes to break the cells. This results in death of the tissue and can impair the function of the organ. The necrotic lesions were characterised by dead cells and nuclear debris. Whereas, early lesions were characterised by the lack of lymphocytes and instead the lymphocytes create an alveolar wall. Inside the alveoli were the accumulation of monocytes or macrophages. What is important to note is that when cutting the early lesions, the alveolar wall was avoided, as it had lymphocytes that contain different proteins not relevant to this project. Therefore, cutting only the cells within the alveolar wall, without the lymphocytes, was crucial and an example of how this was done was with a red line seen in Figure 7.[65-67]

2.5 Proteomics

2.5.1 Sample Preparation

The samples were put through lysis and sonication to release the proteins of interest into the solution. We do this several times until the tissue becomes soft and liquid [68]. To prepare the samples for proteomics, tissue solubilisation and cysteine alkylation is necessary. Alkylating agents will keep the proteins from folding back up again when coated. First, a lysis buffer was created with 4% SDS, 5mM TCEP, 10mM CAA and 0.1 M Tris-HCL pH 8.5. The concentrations were found by using the formula: concentration 1 x volume 1 = concentration 2 x volume 2, and is shown on Table 2.

Table 2: Volume of each chemical needed to create 1000 microliters of lysis buffer.

Chemical	**Formula**	**Volume (microliters)**
4% SDS	20% x X microliters = 4% x 1000 microliters	200
5mM TCEP	50mM x X microliters = 5mM x 1000 microliters	100
10mM CAA	100mM x X microliters = 10mM x 1000 microliters	100
0.1M Tris-HCL pH 8.5 (14 micrograms TRIS + 100 millilitres of H20 + as much HCL to put the pH down to 8.5)	1000 microliters - 400 microliters	600
Total volume of lysis buffer		1000

10 microliters of lysis buffer were added into each sample. The samples were then transferred from the PCR tube into a 1.5 millilitre Eppendorf tube. The samples were heated in a water bath at 98 degrees Celsius with 800 rpm for 1 hour. Heating will stretch the proteins. Next was to sonicate for 30 seconds and allow the samples to rest for 30 seconds, doing this step five times each. Heating in an Eppendorf Shaker and sonication was repeated until tissues were completely homogenised and solubilised and in this project's case, a total of three times of repeating those two steps was needed before centrifuging the samples at 13,000 rpm for 5 minutes. The PEN membrane from LMD-extracted samples will not solubilise and so the supernatant of each sample needs to be transferred into new tubes.

2.5.2 Protein measurement

Next was to determine the protein concentration in the samples by using BCA protein assay, as further into the experiment adjustments must be made for certain concentrations.

Table 3: Solutions needed to be made for BCA protein assay.

Solution	**Ingredients**
200 microlitres BCA mix (50:1 ratio)	9.8mL BCA solution + 0.2mL Copper Sulphate solution

Using a 96 well plate with a flat bottom, water, Bovine IgA/BSA and lysis buffer were added into 6 rows, creating a standard curve to compare the samples to. Into a separate row, 5 microliters of each of the samples was pipetted into the wells along with 10 microliters of water. 200 microliters of BCA mix, a 50:1 ratio by mixing 9.8 millilitres of BCA solution and 0.2 millilitres of Copper Sulphate solution, was added into each well. The plate was covered with parafilm and mixed before being heated at 37 degrees Celsius for 30 minutes in a hot cabinet. As the samples evidently had a high protein concentration, the samples were diluted by splitting the wells into two and adding another 100 microliters of the BCA into each sample well and mixed

and heated again. A diagram of what was put into the wells is displayed on Table 4.

Table 4: Diagram of what was put into each well in microliters.

Well	**1**	**2**	**3**	**4**	**5**	**6**
Standard curve	Water: 10 Bovine IgG/BSA: 0 Lysis buffer: 5 BCA mix: 200	Water: 8 Bovine IgG/BSA: 2 Lysis buffer: 5 BCA mix: 200	Water: 6 Bovine IgG/BSA: 4 Lysis buffer: 5 BCA mix: 200	Water: 4 Bovine IgG/BSA: 6 Lysis buffer: 5 BCA mix: 200	Water: 2 Bovine IgG/BSA: 8 Lysis buffer: 5 BCA mix: 200	Water: 0 Bovine IgG/BSA: 10 Lysis buffer: 5 BCA mix: 200
Sample	Water: 10 Sample: 5 BCA mix: 200	Water: 10 Sample: 5 BCA mix: 200	Water: 10 Sample: 5 BCA mix: 200	Water: 10 Sample: 5 BCA mix: 200	Water: 10 Sample: 5 BCA mix: 200	Water: 10 Sample: 5 BCA mix: 200
Sample diluted	Sample from well 1: 100 BCA mix: 100	Sample from well 2: 100 BCA mix: 100	Sample from well 3: 100 BCA mix: 100	Sample from well 4: 100 BCA mix: 100	Sample from well 5: 100 BCA mix: 100	Sample from well 6: 100 BCA mix: 100

Due to having a low amount of sample, only six samples were looked at and 10 micrograms of protein was assumed to be in the other samples. A plate reader that uses photometry analysed the results. The data was retrieved and is shown on Table 5.

Table 5: Photometry plate reading of BCA Protein Assay, wavelength: 560.

Well	1	2	3	4	5	6
Standard curve	0.859	0.934	1.0523	1.135	1.285	1.421
Samples	0.639	1.010	0.396	1.154	1.159	1.037

2.5.3 Digestion and buffer exchange using SP3

SP3 stands for single-pot, solid-phase enhanced sample preparation, which includes protein reduction and alkylation, protein binding and protein digestion. This process is done for protein clean up. It uses a hydrophilic interaction mechanism to exchange and remove components that are commonly used to facilitate cell or tissue lysis, protein solubilisation and enzymic digestion. Before beginning SP3, the protein amount, sample volume and number of samples must be known. The protein amount is 20 mg/ml, the sample volume is 40 microlitres and the number of samples is 11.

Table 6: Solutions needed for SP3.

Solution	Ingredients
1 mL of 100mM AmBic/1mM CaCl2	Tube 1: 7.9mg AmBic + 990mL H2O Tube 2: 11.09mg CaCl2 + 1,000mL H2O Tube 3: 10 mL into Tube 1
Trypsin 0.2mg/mL	100mL HCL + 20mg trypsin.
Digestion solution: Trypsin/AmBic/CaCl2	52mL trypsin (0.2mg/mL) + 624mL 100mM AmBic/1mM CaCl2
0.5 NaCl	29.2mg NaCl + 1 mL H2O

2.5.3.1 Protein clean up

The first step of SP3 is the protein clean-up to remove the lysis-buffer in exchange for the digestion buffer. To prepare the beads, 31.2 microliters of each bead stock solution was added into a fresh tube. It is important to vortex the beads beforehand. Next, 160 microlitres of water was added into the tube, pipette mixed and placed onto a magnetic rack until the beads have settled onto the tube wall before the supernatant was removed. This step was repeated twice. After, the tube was taken off the magnetic rack and 24.96 microliters of water was added into the tube and pipette mixed. 2 microliters of the beads were added into each sample tube and pipette mixed to homogenise. Then, 121.14 microliters of 100% ethanol was added into each sample to induce binding of the proteins to the beads and pipette mixed. The samples were incubated in a Thermomixer for 7 minutes at 1,000 rpm at 24 degree Celsius. The tubes were placed on the magnetic rack until the beads had settled on the tube wall and the supernatant was removed. Next, 180 microlitres of 80% ethanol was added to each sample and shaken, then put onto the magnetic rack, where the supernatant was removed. This step was repeated a total of four times, due to the low amount of proteins in the sample, it being crucial for rinsing the proteins bound to the SP3 beads and to completely remove traces of SDS.

2.5.3.2 Protein digestion

First, 1 ml of 100mM AmBic/1mM CaCl2 was made by adding 7.9 micrograms of AmBic and 990 microliters of H2O into one tube and in another tube, we added 11.09 micrograms of CaCl2 with 1,000 microlitres of H2O. 10 microliters of CaCl2 into the ambic. To make trypsin concentration of 0.2 micrograms/microlitre, 100 microlitres of HCL was added into 20 micrograms of trypsin to dilute it to 0.2 micrograms/microliters. Next, 52 microliters of trypsin and 624 microliters of Ambic/CaCl2 was added into the same fresh tube to create the digestion solution. 50 microliters of the digestion solution was pipetted into each sample tube. It is important that the sample is not mixed, but that the beads are in the liquid, so, using the pipette, the beads should be pushed into the liquid. After, the samples were sonicated for 30 seconds and incubated for 16 hours at 37 degrees Celsius in a thermomixer at 1,000 rpm.

2.5.3.3 Peptide extraction

The final part of SP3 is the peptide extraction. The sample tubes are centrifuged at 13,000 rpm at 24 degrees Celsius for 3 minutes. The tubes are then placed on the magnetic rack until the beads have settled onto the tube wall before the supernatant was moved to a fresh tube. 0.5M NaCl was prepared by adding 29.2 milligrams of NaCl into 1 millilitre of H2O. The NaCl will make the proteins sticky and so 50 microliters of 0.5 NaCl was added into each sample and pipette mixed. The tubes were then sonicated for 30 seconds in a sonicator and centrifuged at 13,000 rpm at 24 degrees Celsius for 3 minutes. The tubes were placed on the magnetic rack and when the beads settled onto the tube wall, the supernatant of these tubes were moved to the tube with the previous supernatant. The amount should be 100 microliters and combining these two supernatants will release all peptides from the beads. Next, 200 microliters of 10% TFA was added to make the total solution amount 300 microliters. The pH was measured using litmus paper, the pH should be under 5, therefore 5 microliters of 10% TFA was added to each sample.

2.5.4 Oasis peptide desalting

Next is the oasis peptide desalting, which includes column wetting and conditioning and sample blinding, cleaning and elution. This step is crucial as it removes remaining unwanted contaminants and interfering substances. This method is used to de-salt and clean up the sample after protein digestion from various digestion methods.

The oasis 96 well cartridge plate was placed into a waste plate. The cartridges were activated by adding 500 microliters of 80% ACN, 0.1% FA, centrifuge 200 x gram for 1 minute and allowing the discard to flow through. Cartridges were washed by adding 500 microliters of 0.1% TFA, centrifuge 200 x gram for 1 minute and allowing the discard to flow through, and repeated. Next, 500 microliter of the sample was added. If the sample volume is lower than 500 microliters, 0.1% TFA was added to adjust to 500 microliters. It was centrifuged at 100 x gram for 3 minutes and allowed the discard to flow through. To wash, 500 microliters of 0.1% TFA was added, centrifuged 200 x gram for 1 minute and allowed the discard to flow through, this step was repeated two more times. Next, the samples were eluted in the 96 well elution plate with 100 microliters of 80% ACN, 0.1% FA and centrifuged 100 x gram for 3 minutes. This time, the flow

was kept through. The samples were eluted in the 96 well elution plate with 100 microliters of 80% ACN, 0.1% FA and centrifuged 100 x gram for 1 minute. This flow was also kept. The samples were covered with parafilm and freezed at -80 degrees Celsius for fifteen minutes before freeze-dried in a SpeedVac for an hour. When the samples were dried 20 microliters of 1% FA, 2% ACN was added. The samples were then put into the shaker for 5 minutes at 15 rpm and sonicated for 30 seconds.

2.5.5 TMT Labelling

TMT labelling enables users to identify proteins in different samples by chemically labelling them.

Table 7: Solutions needed to be made for TMT labelling.

Solution	**Ingredients**
500 microliters 5% hydroxylamine	50 microlitres 50% hydroxylamine + 450 microlitres 50mM HEPES buffer

To begin TMT labelling, 20 microliters of HEPES buffer were added into each sample before put into a thermomixer for 5 minutes at 1500 rpm. HEPES buffer is used as it does not have any amino acid in it, therefore will not alter the peptides when it is run through the mass spectrometer. The samples were then put onto a nanodrop, and the results can be found on Table 8. Calculations were done to determine how much of each sample was needed using the formula: molecular weight = 1.5 / concentration. The volume of each sample from Table 8 was transferred into a new Eppendorf tube and HEPES buffer was added to make the total concentration 15 microliters. The nanodrop results and the calculations for the molecular weight and HEPES buffer can be found on Table 8.

Table 8: Results from the nanodrop and calculations of molecular weight and amount of HEPES buffer needed.

Sample	Protein (mg/mL)	A280	A260/280	Total protein concentration	HEPES buffer (15 microliters - MW) (microliters)
1	0.663	0.663	1.195	2.262	12.738
2	0.163	0.163	1.383	9.202	5.798
3	0.114	0.114	1.708	13.158	1.842
4	0.318	0.318	1.296	4.717	10.283
5	0.511	0.511	1.344	2.935	12.065
6	0.816	0.816	1.408	1.838	13.162
7	0.406	0.406	1.251	3.695	11.305
8	0.675	0.675	1.149	2.222	12.778
9	0.324	0.324	1.312	4.629	10.371
10	1.688	1.688	1.320	0.889	14.111
11	1.286	1.286	1.129	1.166	13.834

250 microliters of the TMT reagents were added into their perspective tubes. Then, the samples were shaken for 5 minutes at 800 rpm before sonication for 30 seconds and allowing the samples to rest for 5 minutes. 15 microliters of TMT pro reagent were added into each sample and then vortexed at 15000 rpm for 30 seconds. The samples were allowed to rest for 75 minutes. During this time, 500 microliters of 5% hydroxylamine was created by mixing 50 microliters of 50% hydroxylamine and 450 microliters of 50mM of HEPES buffer, refer to Table 7. 5 microliters of the 5% hydroxylamine were added into each sample before vortexed and then incubated for 15 minutes. The samples were then combined into one tube to make 11 micrograms of aliquots. Next, the tube was freezed at -80 degrees for 20 minutes and freeze-dried in a SpeedVac for 1

hour. Finally, 300 microliters of 0.1% TFA was added into the sample tube and shaken at 1500 rpm for 5 minutes and then sonicated for 30 seconds.

2.5.6 Fractionation

Fractionation consists of column conditioning, sample loading, washing and eluting. This part will isolate, identify and characterise various proteins in the sample. To begin, the fractions must be prepared (wash, 2, 4, 6 and 8) and are displayed on Table 9.

Table 9: Preparation of elution solutions for Thermo Scientific TMT-labelled peptides.

Fraction	Acetonitrile (%)	Acetonitrile (microliters)	0.1% Triethylamine (microliters)
Wash	5.0	50	950
1	12.5	125	875
2	17.5	175	825
3	22.5	225	775
4	50.0	500	500

The white tip of the column was removed and the column was placed into a 2 mililiter Eppendorf tube before centrifuged at 5000 x grams for 2 minutes. This will remove the solution and pack the resin material. The liquid was discarded. 300 microliters of ACN was loaded into the column, centrifuged at 5000 x g for 2 minutes and the ACN was discarded. This step was repeated once more. The column was then placed into a new tube and 300 microliters of the sample solution was pipetted into the column, before being centrifuged at 3,000 x g for 2 minutes. The tube is kept and the column was placed into a new tube. This step was repeated for each of the fractions, refer to Table 9. Next, the tubes were freezed at -80 degrees for 20 minutes and freeze-dried until the liquid had completely evaporated. 8 microliters of 0.1% formic acid was added to sample 4 and 7 microliters of formic acid was added into sample 1, 2 and 3. This is so that the samples

have roughly 1 microgram of protein and enough to put 1 microliter into the nanodrop. 6 microliters is the max volume able to be put into the mass spectrometer, so that if there is only a small amount of protein found in the samples, all of it can be put into the mass spectrometer. The samples were then shaken at 1500 rpm for 5 minutes and sonicated for 30 seconds before using 1.5 microliters of each of the samples to run on a nanodrop. The results of the nanodrop are displayed on Table 10.

Table 10: Protein concentration in each of the 4 fractions using nanodrop.

Fraction	**Protein (mg/mL)**	**A280**	**A260/280**
1	0.553	0.553	1.334
2	0.374	0.374	1.745
3	0.505	0.505	1.631
4	0.544	0.544	1.644

2.5.7 Mass Spectrometer

Finally, we put the sample into the mass spectrometer which is a machine used to measure the mass to charge ratio of the molecules present in the sample at certain points in time.

The settings must first be altered and details of what was used for this project can be found in Appendix 1. First, a blank was inserted into the mass spectrometer in order to try and get rid of contamination from previous users. Next, a Helia sample was run through the mass spectrometer, which acts as a control for the HBLC and the mass spectrometer in order to check that everything is functioning correctly. This is then followed by another blank, before finally allowing the real samples to be run through. 1.3 microliters was taken from each sample except for 2, because it had a protein concentration of 0.2 and so more was required to compensate for this and 2.3 microliters was inserted into the mass spectrometer.

2.6 Data analysis

After retrieving the results from the mass spectrometer, analysing the data was the next step.

2.6.1 Creating a lung specific microbiome FASTA file

A FASTA file is a text-based document containing amino acid sequences. There were a total of 226,274 entries across 25 species in the database and can be found on Table 11.

Table 11: Database search list that includes the name of the bacteria, the number of proteomes in the database, the code of the proteome database and what the bacteria is associated with.

Name	Filter	Code	Proteomes	Reason for selection
Homo sapiens	Human	UP000005640	79038	Human
Mycobacterium tuberculosis	MYCTU	UP000001584	3993	TB
Clostridium tetani	CLOTE	UP000001412	2415	
Bacteroides	BACTN	UP000001414	4782	Decrease in bacteroides associated with TB antibiotic treatment
	BACFN	UP000006731	5518	
Prevotella	PRER2	UP000000927	2761	TB, COPD and increase in SCFA
Helicobacter pylori	HELPY	UP000000429	1554	If found in the gut, patients were 50% less likely to develop TB
	HELHP	UP000002495	1873	
Klebsiella pneumoniae	KLEP7	UP000000265	5127	Pneumonia
	KLEP3	UP000001734	5738	
	KLEPH	UP000007841	5728	
Streptococcus pneumoniae	STRR6	UP000000586	2030	
	STRPN	UP000000585	2115	
	STRP2	UP000001452	1915	
	STRPS	UP000001682	2194	

Pneumococcus	BPDP1	UP000008920	72	
Staphylococcus aureus	STAA8	UP000008816	2889	
	STAAN	UP000000751	2580	
	STAAW	UP000000418	2660	
	STAAE	UP000006386	2583	
Moraxella catarrhalis	MORCB	UP000000930	1881	
Escherichia coli	ECOL6	UP000001410	5336	
	ECO24	UP000001122	4915	
	ECO27	UP000008205	4594	
Enterobacteriaceae	ENTBF	UP000011002	4320	
Listeria monocytogenes	LISMO	UP000000817	2844	Pulmonary infections
	LISM4	UP000001288	2815	
	LISMG	UP000016703	2847	
Achromobacter xylosoxidans	ACHXA	UP000006876	6853	Cystic fibrosis
Haemophilus influenzae	HAEIN	UP000000579	1708	
Burkholderia cenocepacia	BURCJ	UP000001035	6993	
Pseudomonas aeruginosa	PSEAE	UP000002438	5564	Cystic fibrosis, bronchopulmonary dysplasia and COPD
	PSEA7	UP000001582	5975	
Mycobacterium avium	MYCA1	UP000001574	5054	Complex lung disease
Ureaplasma parvum	UREPA	UP000000423	611	Chronic lung disease
	UREP2	UP000002162	609	
Veillonella	VEIPT	UP000007968	1846	HIV
Candida albicans	YEAST	UP000000559	6035	Allergies
Bifidobacterium infantis	BIFL1	UP000008150	1975	
	BIFLS	UP000001361	2399	
Neosartorya fumigatus	ASPFU	UP000002530	9647	
Neisseria	NEIMA	UP000000626	1887	Air pollution

	NEIMB	UP000000425	2001	

First, the *homo sapiens*, contaminants and MTB proteome database were added as these are the three obvious subjects to look at. Based on previous peer-reviewed published literature [69-74], other databases related to TB and the lungs were combined and searched. These bacterias include being associated with pneumonia, pulmonary diseases and infections, HIV, allergies and air pollution. Using Uniprot [75], the proteomes were downloaded and then combined to create a database.

2.6.2 Bioinformatics analysis

Bioinformatics uses software programs that assists in understanding and cleaning, for example normalising the results, especially for large biological data.

2.6.2.1 Mapping of MS/MS spectra to our database using Thermo Proteome Discoverer version 2.5

Thermo Proteome Discoverer version 2.5 [76] combines the database and the raw data from the mass spectrometer. It is also able to normalise the data. Figure 8: a shows the processing method which matches the spectra to the database, uses false discovery rate (FDR) and finds the confidence of the peptides identified. Figure 8: b shows the consensus method which quantifies the data, particularly comparing the between samples and normalisation of the total peptide amount. See Appendix 2 and 3 for more details on the settings.

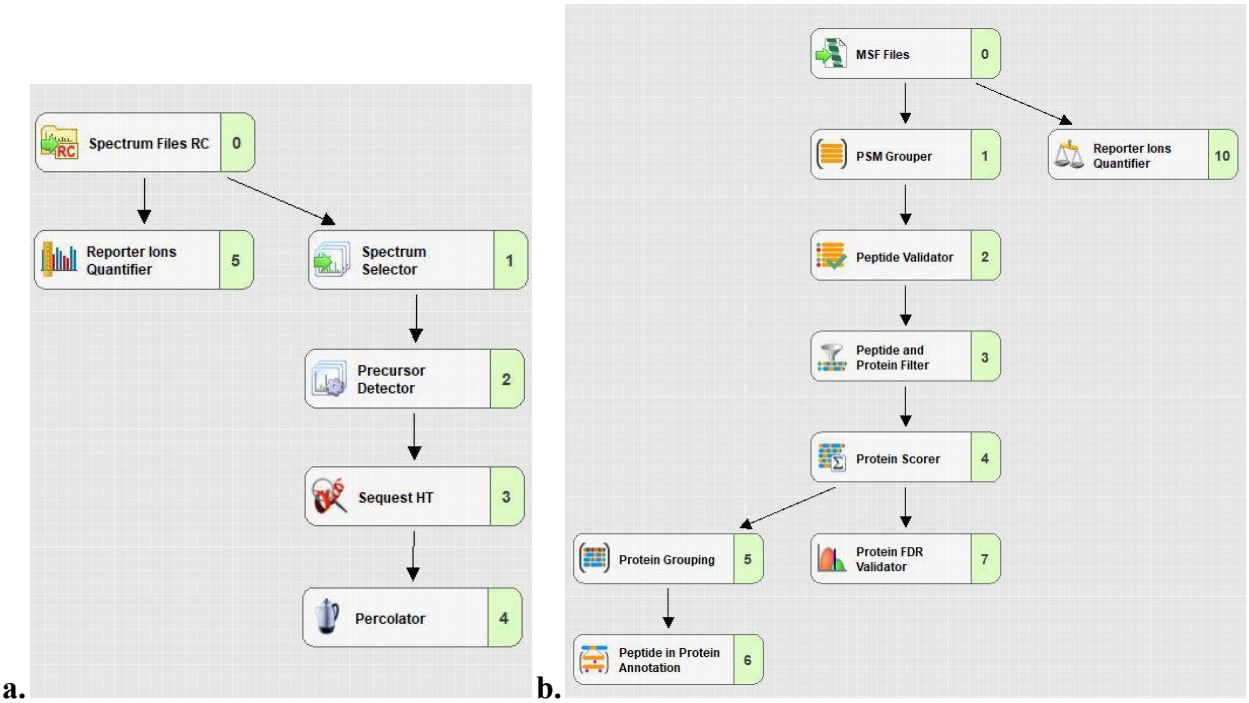


Figure 8: Flow chart of settings used for Thermo Proteome Discoverer. **a |** Processing method. **b |** Consensus method.

2.6.2.2 Quality checking and statistical analysis using Perseus version 1.6.15.0

The data was then transferred to Perseus version 1.6.15.0 [77] for quality checking and statistical analysis. With Perseus, histograms, volcano plots and heatmaps are able to be created. Figure 9 shows the steps done in Perseus to get each of the diagrams that can be found in the results section.

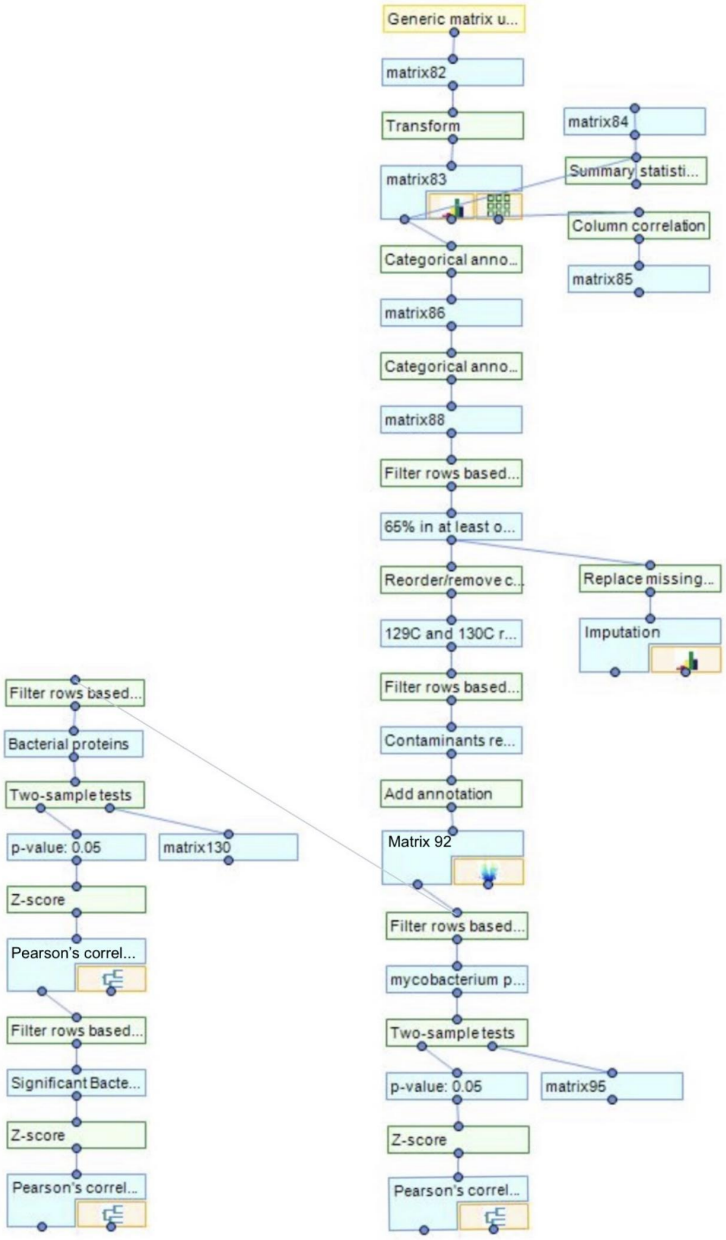


Figure 9: Process of data in Perseus version 1.6.15.0.

First, the data was log2 transformed. Next, summary statistics, a set of simple descriptive quantities that are calculated and summarise the data, and column correlation, calculating correlation coefficients between the columns, was done, along with creating histograms and a multi-scatter plot to assist in visualising the overall data. Moving on, the data was categorically

annotated, so each sample was labelled with their prospective patient case number and labelled again according to lesions; early lesions and necrotic lesions. Then, the rows of the data were filtered by only valid values of 65% in at least one group.

Imputation was attempted to see if the data would improve, however was discarded and will be discussed more in the results and discussion section as to the reason for this. So, the data continued by removing two of the samples, 129C and 130C, which again will be discussed further on. The rows were then filtered, and contaminants were removed from further results. Then, annotations were added; GOBP name, GOMF name, GOCC name, GOBP slim name, GOCC slim name, KEGG, KEGG name, Pfam, GSEA, Keywords and Gene name. Using this data, a volcano plot was created to visualise the results and which proteins were considered statistically significant.

The rows were again filtered so that all proteins that were categorised under mycobacterial proteins were kept and those that did not were removed. A two-sample test using a p-value of 0.05 was conducted before filtering out the *homo sapiens* proteins and then doing the same process as with the *homo sapiens* proteins. The two sample tests with a p-value of 0.05 was more suited for the bacteria proteins as there were much less proteins found compared to the FDR permutation. A z-score, the mean of the data subtracted from each value and then divided by the standard deviation, was immediately done on the data using Pearson's correlation to visualise all the mycobacterial proteins on a hierarchical clustering.

The same steps done for the mycobacterial proteins were done for all bacterial proteins, however only *homo sapiens* proteins were removed and after the first hierarchical clustering, a second one was made with only significant bacterial proteins.

2.6.2.3 Excel

Excel was used to create line graphs to illustrate a comparison of protein abundance in each sample and between lesions.

2.6.2.4 Calculations and graphs

2.6.2.4.1 Histograms

Histograms are created on a common scale; therefore, it provides an easy way to compare and analyse distributions between and within the samples.

2.6.2.4.2 Volcano plot

Volcano plots measure statistical significance using -log p-value on the y-axis and magnitude of change using Z-score on the x-axis. Each point represents a protein that has been found in the samples. Proteins that have been identified above the FDR cut-off at a significance level of 1.2 are considered to be statistically significant.

2.6.2.4.3 Hierarchical clustering

Hierarchical clustering enables a visualisation of trends and issues among the different samples and proteins. There are two types of hierarchical clustering; unsupervised and supervised. Unsupervised is a more general display of the samples and proteins, as opposed to supervised where proteins are specifically selected to be looked at.

Pearson's correlation was decided over Euclidean distance because Pearson's is more commonly used among researchers to study gene expression, it measures shape between proteins and considers the overall expression of proteins. Also, when using the standard Euclidean distance and comparing the hierarchical clustering to the one created with Pearson's correlation, the Pearson's correlation clustered the data better than the Euclidean distance.

2.6.2.4.4 Exponentially modified protein abundance index

An established method to estimate absolute protein concentration from peptide counts in proteomics is called exponentially modified protein abundance index (emPAI) where it creates a ratio of observed to observable peptides[78]. The equation to determine the emPAI of an identified protein is as follows:

$$emPAI = 10^{\frac{Number\ of\ observed\ peptides}{Number\ of\ observable\ peptides}} - 1$$

The number of observed peptides was determined as "# of peptides" from Thermo Proteome Discoverer as this is the total number of distinct peptide sequences identified for the protein. The number of observable peptides was determined as "# of PSMs" from Thermo Proteome Discoverer as it is the total number of peptide-spectrum matches identified from all included searches.

2.7 Protein interaction and network analysis

Protein interaction and network analysis is the exploration and examination of the relationship and interaction between two or more proteins. String Database version 11.5 [79] and Cytoscape version 3.9.1 [80] were used to show the protein interaction. String was used first to enter the proteins and then transferred to Cytoscape for creating readable diagrams that include fold change.

2.8 Comparing results with the histology in conjunction with medical records

A comparison of the results from the proteomics with the histology in conjunction with medical records was done.

3 Results

3.1 Description of patients

Displayed on Table 12 are the characteristics of each patient used in this study. There were five participants with a total of eleven samples extracted; six early lesions and five necrotic lesions. The range of death was between 1935 – 1947 with 80% of the patients being male. All had pulmonary TB and patients 2, 3 and 5 also had extrapulmonary and miliary TB. All patients presented at least one symptom of TB and a cavity in the apex of the lung was found in patient 3. 80% of the patients' death was caused by TB and of these only one was due to pulmonary TB. Patient 3 had meningitis, where the membranes around the brain and spinal cord become inflamed which can be caused by TB and is called tuberculous meningitis [81].

Table 12: Patient characteristics

Patient	**1**	**2**	**3**	**4**	**5**
Year of death	1946	1947	1935	1943	1936
Gender	Male	Male	Male	Female	Male
Age	38 years	45 years	44 years	5 months	20 years
Type of TB	Pulmonary	Pulmonary, extrapulmonary, miliary	Pulmonary	Pulmonary, extrapulmonary, miliary	Pulmonary, extrapulmonary, miliary
Symptoms	Symptoms of ulcer	Fever	Cough with yellowish expectorate, fever, night sweats	Cough, fever, weight loss, cavities and TB +ve on expectorate	Headache
TB testing	N.A.	Pleural exudate culture +ve	N.A.	Sputum +ve, tuberculin +ve	Culture +ve from both lung and liver
Cavities	No	No	No	Apex of lung	No
Death caused by TB	No, death due to ulcer perforated	Yes	Yes	Yes	Yes
Death caused by	N.A.	No	Yes	No	No

pulmonary TB					
Meningitis	No	No	No	Yes	No
Lesions taken	2 early + 1 necrotic	2 early + 1 necrotic	2 early + 1 necrotic	1 necrotic	1 necrotic

Abbreviations: N.A., not applicable; +ve, positive.

3.2 Normalising data

3.2.1 Protein abundance

Protein abundance is defined as the amount of proteins that have been identified in the sample based on the database that was used.

Table 13: The total amount of identified proteins found in each sample.

Sample	Lesion	Total number of identified proteins
1.1	Necrotic	3675
1.2	Early	3674
1.3	Early	3638
2.1	Necrotic	3676
2.2	Early	2057
2.3	Early	3671
3.1	Necrotic	2644
3.2	Early	3424

3.3	Early	3675
4.1	Necrotic	3676
5.1	Necrotic	3675

The numbers are quite similar except for 2.2 and 3.1, which are significantly lower and potential outliers compared to the rest of the samples.

3.2.2 Quality measurements and normalisation of evaluation

Histograms of each of the samples provides an easy to visualise and analyse comparison both within and between distributions as they are created on a common scale.

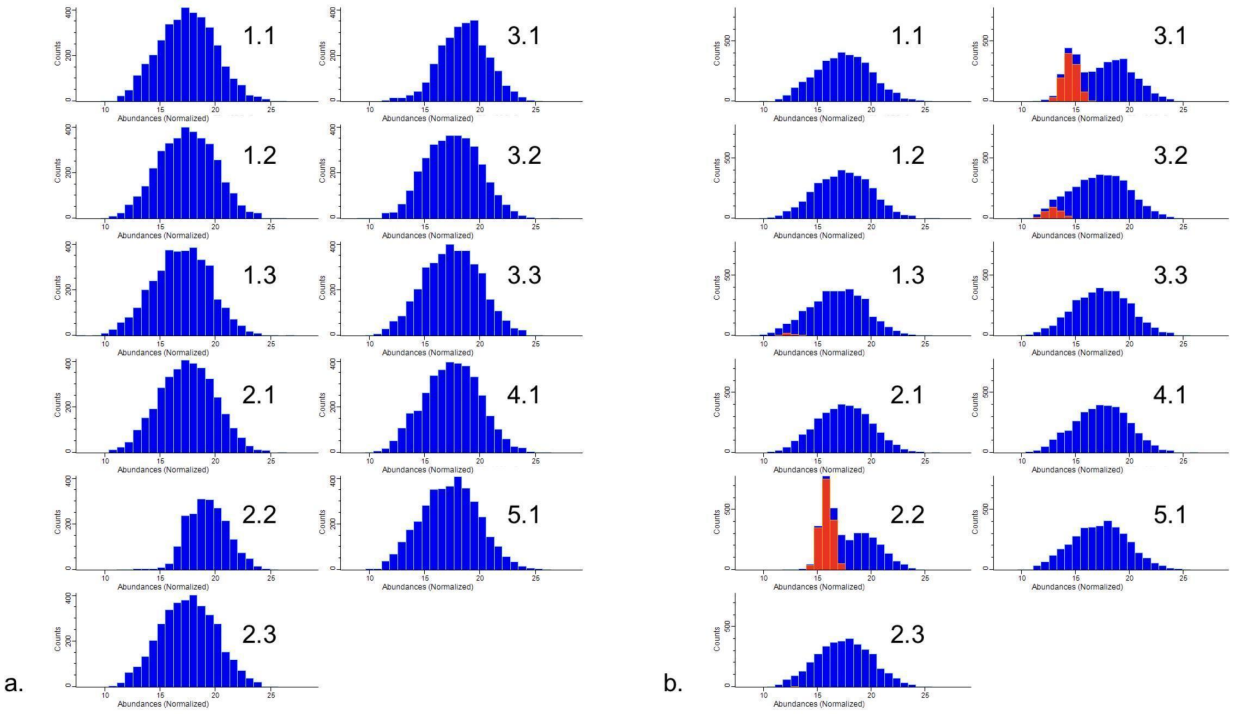


Figure 10: Histogram of abundance of proteins in all eleven samples from all five cases. Graphical representation of the plots were generated by Perseus software version 1.2.15.0. **a |** non-imputed samples. **b |** imputed samples.

All samples were shown to have a similar abundance distribution, except for sample 2.2 and 3.1. Both samples were skewed slightly to the right and do not have enough proteins found, shown in Figure 10: a. The same samples also had the lowest number of identified proteins. With the use of statistics, missing data were replaced with values by imputation, displayed on Figure 10: b. The red bars show the values that were replaced. Samples 2.2 and 3.1 have an increased number of proteins that were added to the sample and still do not look normally distributed. Removing the proteins from the other samples was not a possibility as this would have discarded around 56% of the data. Therefore, samples 2.2 and 3.1 were considered outliers and hence, removed from further analysis. All results were conducted without imputation.

3.3 Bacterial proteins

3.3.1 Protein abundance

Table 14 lists the different bacterial species and the *homo sapiens* proteins identified from the samples. A total of 3531 *homo sapiens* proteins were identified and 110 bacterial proteins identified. Of these bacterial proteins, 5 were mycobacterial proteins; 1 *Mycobacterium avium* and 4 *Mycobacterium tuberculosis*.

Table 14: Table of bacteria proteins with their identified protein amount. Mycobacterial proteins are highlighted in yellow.

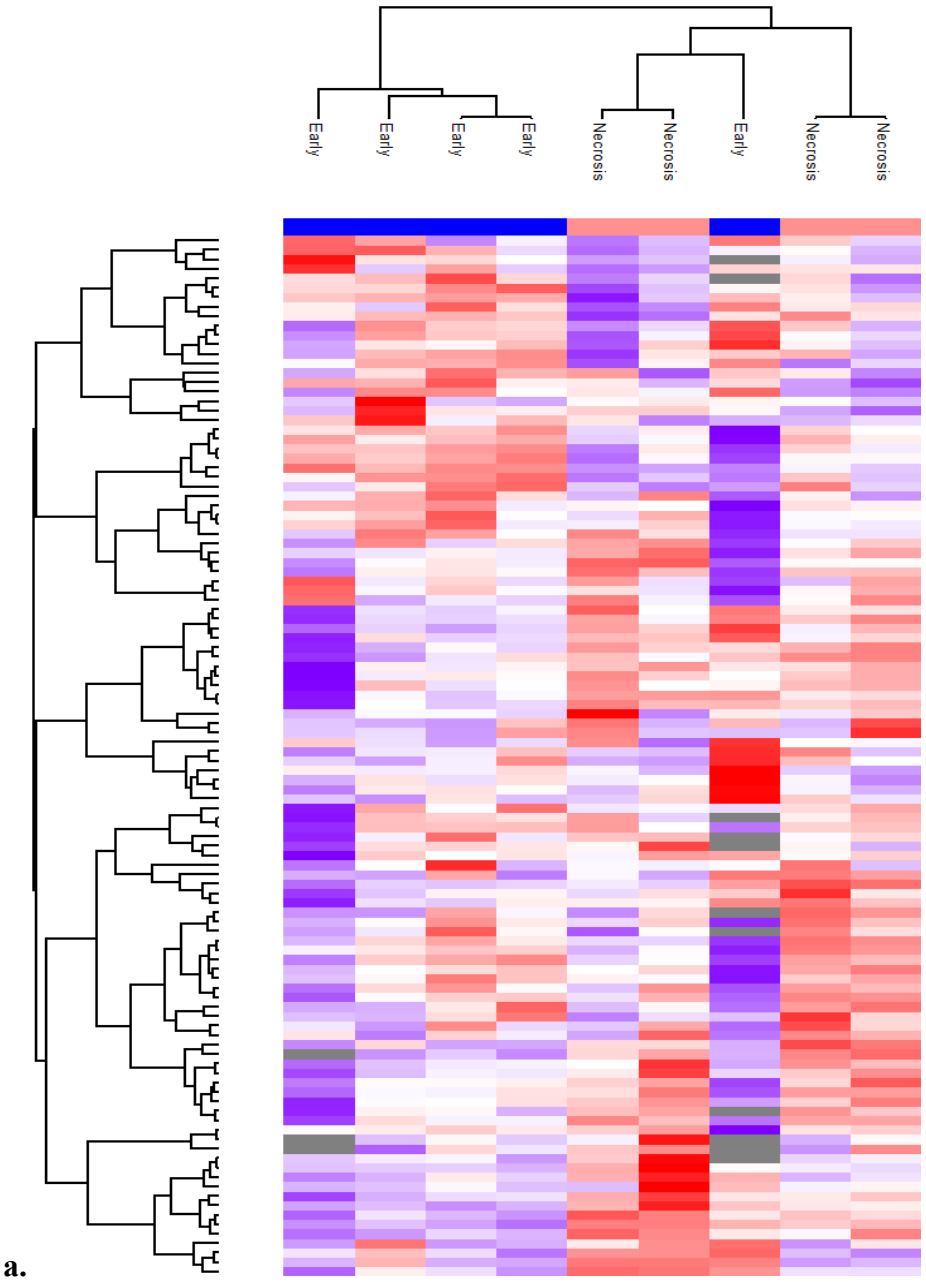
Bacteria	**Number of identified proteins**
Neosartorya fumigata	19
Candida albicans	14
Burkholderia cenocepacia	7
Escherichia coli	7

Pseudomonas aeruginosa	7
Achromobacter xylosoxidans	6
Clostridium tetani	4
Klebsiella pneumoniae	4
Mycobacterium tuberculosis	4
Helicobacter pylori	3
Listeria monocytogenes serotype 1/2a	3
Neisseria meningitidis serogroup A / serotype 4A	3
Neisseria meningitidis serogroup B	3
Staphylococcus aureus	3
Veillonella parvula	3
Bacteroides fragilis	2
Bacteroides thetaiotaomicron	2
Bifidobacterium longum subsp. Infantis	2
Haemophilus influenzae	2
Moraxella catarrhalis	2
Prevotella ruminicola	2
Streptococcus pneumoniae	2
Streptococcus pneumoniae serotype 4	2

Enterobacteriaceae bacterium	1
Helicobacter hepaticus	1
Mycobacterium avium	1
Ureaplasma parvum serovar 3	1

3.3.2 Hierarchical clustering of bacterial proteins

Hierarchical clustering, using Pearson's correlation, was used to see if the TB samples could be separated based on bacterial protein expression. It enables a visualisation of trends and issues among the different samples and proteins. The contaminants have been removed in the following hierarchical clustering.


Early
Early
Early
Early
Necrosis
Necrosis
Early
Necrosis
Necrosis

a.

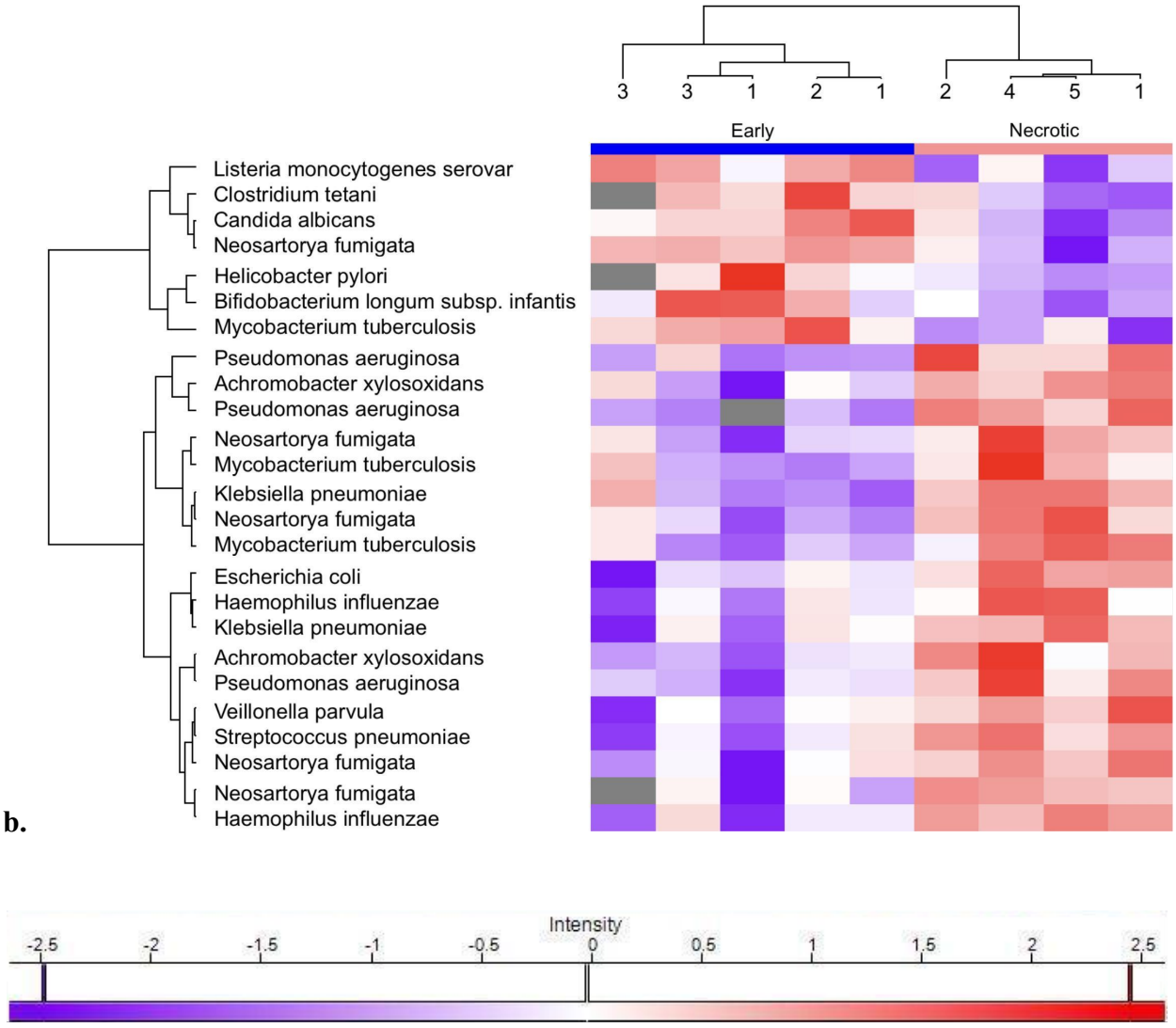


Figure 11: Hierarchical clustering of significantly differentially expressed bacterial proteins in TB samples. It displays the log2 intensity z-score of the eleven TB samples. The bar chart represents the z-score. Graphical representation of the interaction was generated by Perseus software version 1.2.15.0. **a |** Unsupervised hierarchical clustering of all bacteria proteins in each of the eleven samples. **b |** Supervised hierarchical clustering of significantly differentially expressed bacteria proteins in each of the eleven samples.

Figure 11: a is an unsupervised hierarchical clustering as the proteins were not specifically selected to be looked at and instead all bacterial proteins were displayed. There are a total of 110

proteins with 25 significant proteins. The necrotic and early lesions are apart from one another, which shows that the proteins found are clearly expressed differently. However, there is one early lesion sample that has been grouped between necrotic lesions. Figure 11: b is a supervised hierarchical clustering as the proteins were specifically selected to be looked at. Along the left side are the species of bacteria of each of the significantly differentially expressed bacterial proteins. There are a total of 25 significant bacterial proteins.

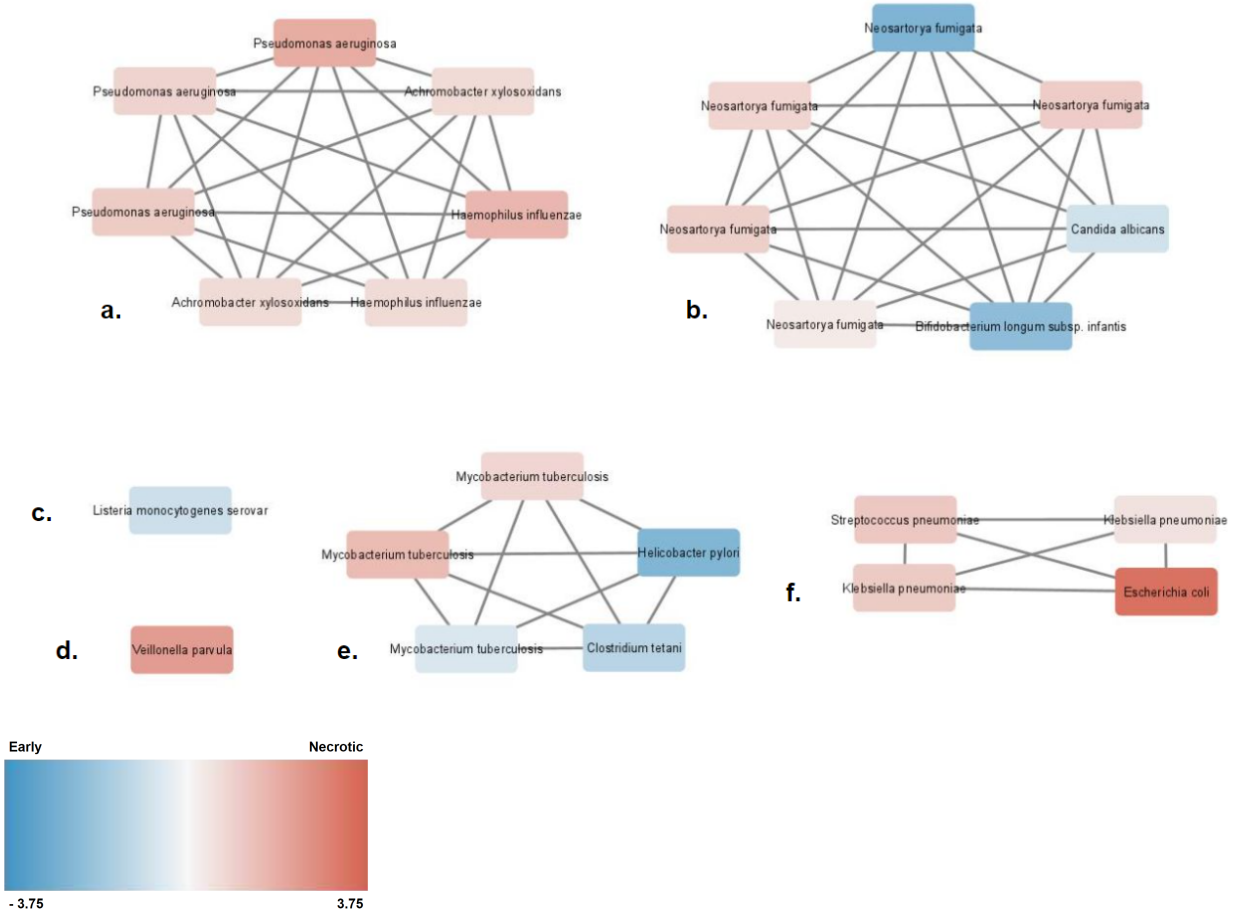


Figure 12: Significantly differentially expressed bacterial proteins found in all the samples with fold change, grouped by associated causation of infection or disease. The bar chart represents the fold change. Graphical representation of the interaction was generated by Cytoscape software version 3.9.1. **a** | Cystic fibrosis. **b** | Allergies. **c** | Pulmonary infections. **d** | HIV. **e** | Tuberculosis. **f** | Pneumonia.

The bacteria in Figure 12 are grouped by what infection or disease they are associated with according to previous published literature, refer to Table 11. All bacterial proteins associated with cystic fibrosis, HIV and pneumonia have a higher interaction with the necrotic lesions as compared to the early lesions. The protein from *Listeria monocytogenes serovar* associated with pulmonary infections interacts more in the early lesions. Allergies and TB have a mix of bacterial proteins interacting in both early and necrotic lesions. Below, Table 15 displays a list of all the significant bacteria proteins found in the lesions. Listed are their accession key, gene name, bacteria, a brief description of the proteins and their log2 fold change. When the fold change is positive, it interacts more with the necrotic lesions, whereas in the negatives, the protein interacts more in the early lesions.

Table 15: Significantly differentially expressed bacteria proteins.

Accession	**Gene**	**Bacteria**	**Description**	**Log2 fold change (necrotic - early)**
A0A0H2V898	c2405	*Escherichia coli*	Uncharacterised protein	3.138
A0A1D8PQZ0	CRN1	*Candida albicans*	Coronin	-0.765
A4D9K1	AFUA_1G08795	*Neosartorya fumigata*	Shugoshin family protein	0.808
A6TCD1	yfgL	*Klebsiella pneumoniae*	Outer membrane protein assembly factor BamB	0.505
A6TCR1	KPN_02976	*Klebsiella pneumoniae*	Putative bacteria extracellular solute-binding protein,	1.046

			family 3	
A6V4X5	PSPA7_2750	*Pseudomonas aeruginosa*	Putative transcriptional regulator	1.776
B7GUL0	Blon_2093	*Bifidobacterium longum subsp. infantis*	DUF4365 domain-containing protein	-2.066
D1BM10	Vpar_0736	*Veillonella parvula*	SMI1_KNR4 domain-containing protein	2.142
E3HFI1	AXYL_01937	*Achromobacter xylosoxidans*	Uncharacterised protein	0.661
E3HMD8	AXYL_03696	*Achromobacter xylosoxidans*	ABC transporter family protein	0.670
O25668	HP_1024	*Helicobacter pylori*	Co-chaperone-curved DNA binding protein A (CbpA)	-2.334
O33203	cycA	*Mycobacterium tuberculosis*	Probable D-Serine/alanine/glycine transporter protein CycA	0.778
P44304	gapA	*Haemophilus influenzae*	Glyceraldehyde-3-phosphate dehydrogenase	0.667
P44909	pdxH	*Haemophilus influenzae*	Pyridoxine/pyridoxamine 5'-phosphate oxidase	1.509
P71658	mihF	*Mycobacterium tuberculosis*	Purative interrogation host factor MihF	1.401

P9WQA5	Rv2971	*Mycobacterium tuberculosis*	Uncharacterised oxidoreductase Rv2971	-0.581
Q4WHG5	lia1	*Neosartorya fumigata*	Deoxyhypusine hydroxylase	0.933
Q4WJ30	AFUA_1G07440	*Neosartorya fumigata*	Molecular chaperone Hsp70	0.287
Q4WWR1	AFUA_3G06760	*Neosartorya fumigata*	60S ribosomal protein L37	-2.375
Q4X0D4	AFUA_2G13860	*Neosartorya fumigata*	Histone H4	1.009
Q898G4	CTC_00488	*Clostridium tetani*	Conserved protein	-1.229
Q8YA77	lmo0282	*Listeria monocytogenes serovar*	Lmo0282 protein	-0.854
Q97NX6	scpB	*Streptococcus pneumoniae*	Segregation and condensation protein B	1.120
Q9HZI6	PA3020	*Pseudomonas aeruginosa*	Probable soluble lytic transglycosylase	1.026
Q9I479	PA1265	*Pseudomonas aeruginosa*	Uncharacterised protein	0.860

3.4 Certainty of mycobacterial proteins identifications

3904 proteins from the bacteria and human database were identified. 3904 protein groups, 22050 peptide groups and 57058 peptide-spectrum matches (PSMs) were identified.

3.4.1 Statistical strength

Table 16 shows the more detailed data of the identified bacterial proteins. As the samples were run through the mass spectrometer twice, there are multiple data for three of the proteins.

Table 16: Raw data from Thermo Protein Discoverer version x.

Accession	**Gene**	**Confidence**	**PSM ambiguity**	**Number of peptides**	**Isolation interference (%)**	**Percolator q-value**	**Percolator PEP**
P71658	mihF	High	Unambiguous	1*	86	0.006	0.047
P9WPE7	groEL2	High	Unambiguous	2*	19	0.000039	0.00011
P9WPE7	groEL2	High	Unambiguous	2*	11	0.00011	0.00035
P9WPE7	groEL2	High	Unambiguous	2*	21	0.00011	0.00058
A0A0H3A054	hup	High	Unambiguous	1*	36	0.0011	0.011
A0A0H3A054	hup	High	Unambiguous	1*	49	0.0023	0.022
P9WQA5	Rv2971	High	Unambiguous	1*	20	0.001	0.01
P9WQA5	Rv2971	High	Unambiguous	1*	12	0.0014	0.015
P9WQA5	Rv2971	High	Unambiguous	1*	31	0.0035	0.03
O33203	cycA	High	Unambiguous	1*	8	0.0014	0.015

Abbreviations: PSM, peptide spectrum matches; PEP, posterior error probability; *, unique peptides.

All proteins were found to have a high confidence and the PSM is the only match as there is no ambiguity. These proteins also have a total of 1 - 2 peptide sequences identified in the protein group and each of those peptides were unique to the particular protein. The isolation interference is quite low in all the proteins except for mihF and hup, which have a percentage of 86 and average of 42.5 respectively. Due to TMT labelling and certain proteins, for example mihF and hup, having similarities with others, the interference was raised. Percolator increases yield and assigns reliable confidence measures such as q-value and PEP. Q-value finds the false positives in the data and PEP finds the probability that the observed PSM is incorrect and is instead random. Both statistics have found that all proteins had a q-value of less than 0.006 and a PEP of less than 0.04, considering the data highly confident with a very low chance it is incorrect.

3.4.2 Manual interrogation of mass spectra from bacterial proteins

MS1 shows all the potential peptide matches. Highlighted in yellow is the peptide Proteome Discoverer deemed as the best match using statistical methods. This peptide is then fragmented and the mass of the peptide fragments is measured in MS2.

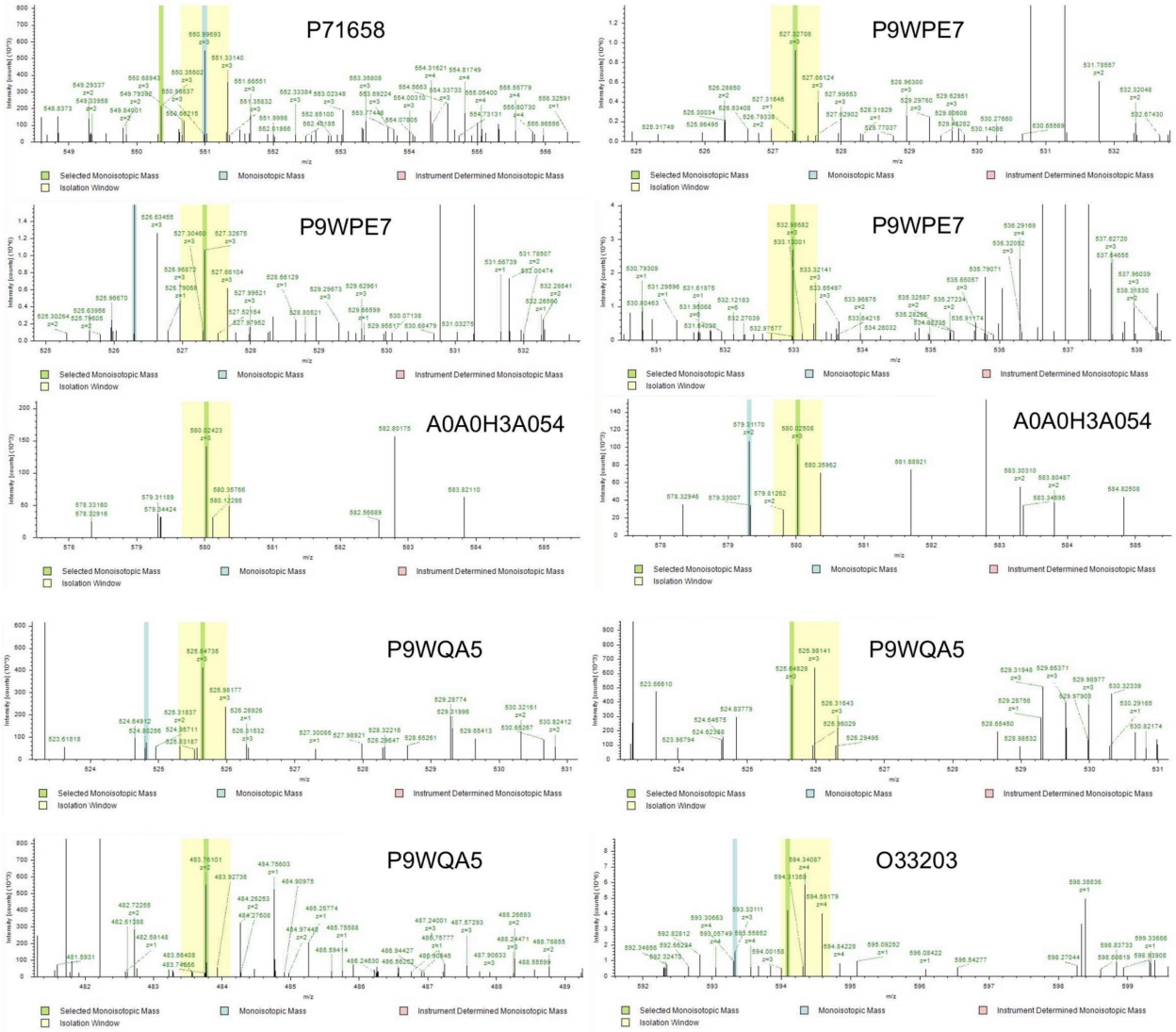


Figure 13: Precursor isotope pattern of each sample, MS1. The y-axis shows the intensity counts at 10^3 and the x-axis shows mass-to-charge ratio as m/z values. Highlighted in yellow is the protein selected to be identified and further fragmented in MS2. Graphical representation of MS1 was generated by Thermo Proteome Discoverer software version 2.5.

Figure 13 displays the precursor isotope pattern from each sample. All samples have identified a single peptide. P71658 was found to have the highest percentage of interference shown in Table 16 and this is evident as many peptides were found to be potential matches shown on Figure 13, however, Proteome Discoverer managed to define a single peptide.

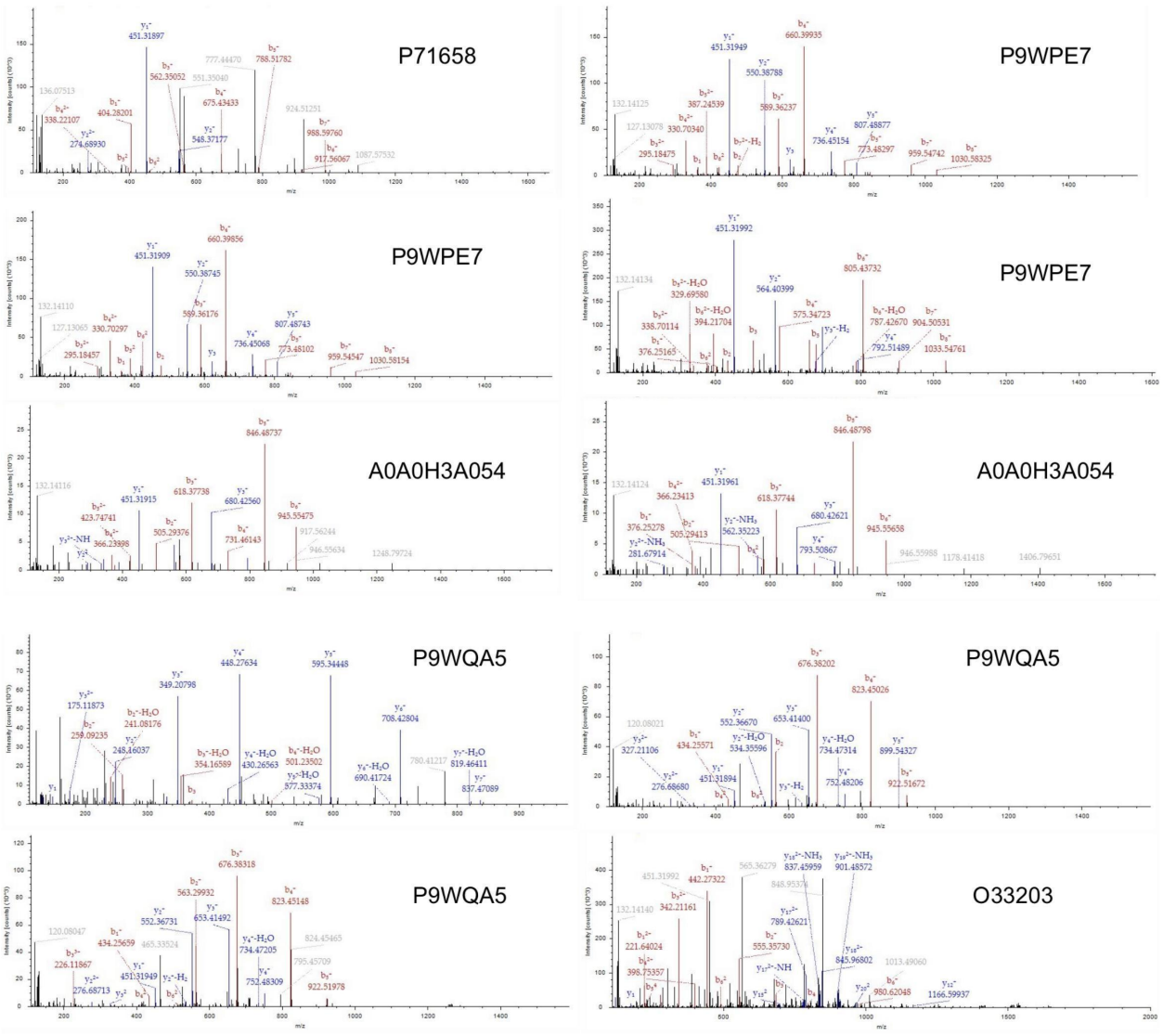


Figure 14: Fragment mass spectrum of each sample, MS2. The y-axis shows the intensity counts at 10^3 and the x-axis shows mass-to-charge ratio as m/z values. The y- and b-ions are annotated in blue and red, respectively, and mass values not used for identifications were not annotated. Graphical representation of MS2 was generated by Thermo Proteome Discoverer software version 2.5.

Figure 14 displays the fragment mass spectrum of each sample. MS2 relates to MS1 as it is a closer look at MS1, the peptide that was highlighted in yellow from Figure 13 was fragmented

and displayed on Figure 14. Both y- and b-ions were abundantly found in each of the samples, meaning that the peptides have been strongly identified. P71658 and A0AH3A054 both had the highest percentage of interference, refer to Table 16, compared to the other samples and it shows in Figure 14 as not as many ions were identified in these samples.

3.4.3 Exponentially modified protein abundance index

emPAI creates a ratio of observed to observable peptides, in order to provide approximate, label-free, quantification of abundance of proteins in the samples[78]. Calculations was done to determine each protein's emPAI value and is shown below in Table 17.

Table 17: Calculations of the emPAI of the mycobacterial proteins from this project

Accession	**Gene**	**Number of observed peptide**	**Number of observable peptides**	**emPAI**
P71658	mihF	1	1	9.000
P9WPE7	groEL2	2	3	3.642
A0A0H3A054	hup	1	2	2.162
P9WQA5	Rv2971	1	3	1.154
O33203	cycA	1	1	9.000

Table 18, compares the emPai score calculated in Table 17 with other studies investigating mycobacterial proteins. These studies, however, worked with culture TB rather than actual samples and some had not displayed all of their emPAI scores and hence not all mycobacterial proteins were found with an emPAI score to compare to.

Table 18: Comparing the emPAI calculated from this project with other studies.

Order lotus name	Gene	emPAI from this project	[82]	[83]	[84]	[85]	[86]
Rv1388	mihF	9.000	4.78	0.6268	N.A.	N.A.	N.A.
Rv0440	groEL2	3.642	55.23	N.A.	0.0626	3.20	0.76
MAV_3836	hup	2.162	N.A.	N.A.	N.A.	N.A.	5.09
	Rv2971	1.154	4.99	N.A.	N.A.	N.A.	N.A.
Rv1704c	cycA	9.000	N.A.	N.A.	N.A.	N.A.	N.A.

Abbreviations: N.A., not available.

3.5 Comparison of mycobacterial proteins in early and necrotic lesions

3.5.1 Volcano plot

Volcano plots measure statistical significance using -log p-value on the y-axis and magnitude of change using Z-score on the x-axis. Each point represents a protein that has been found in the samples.

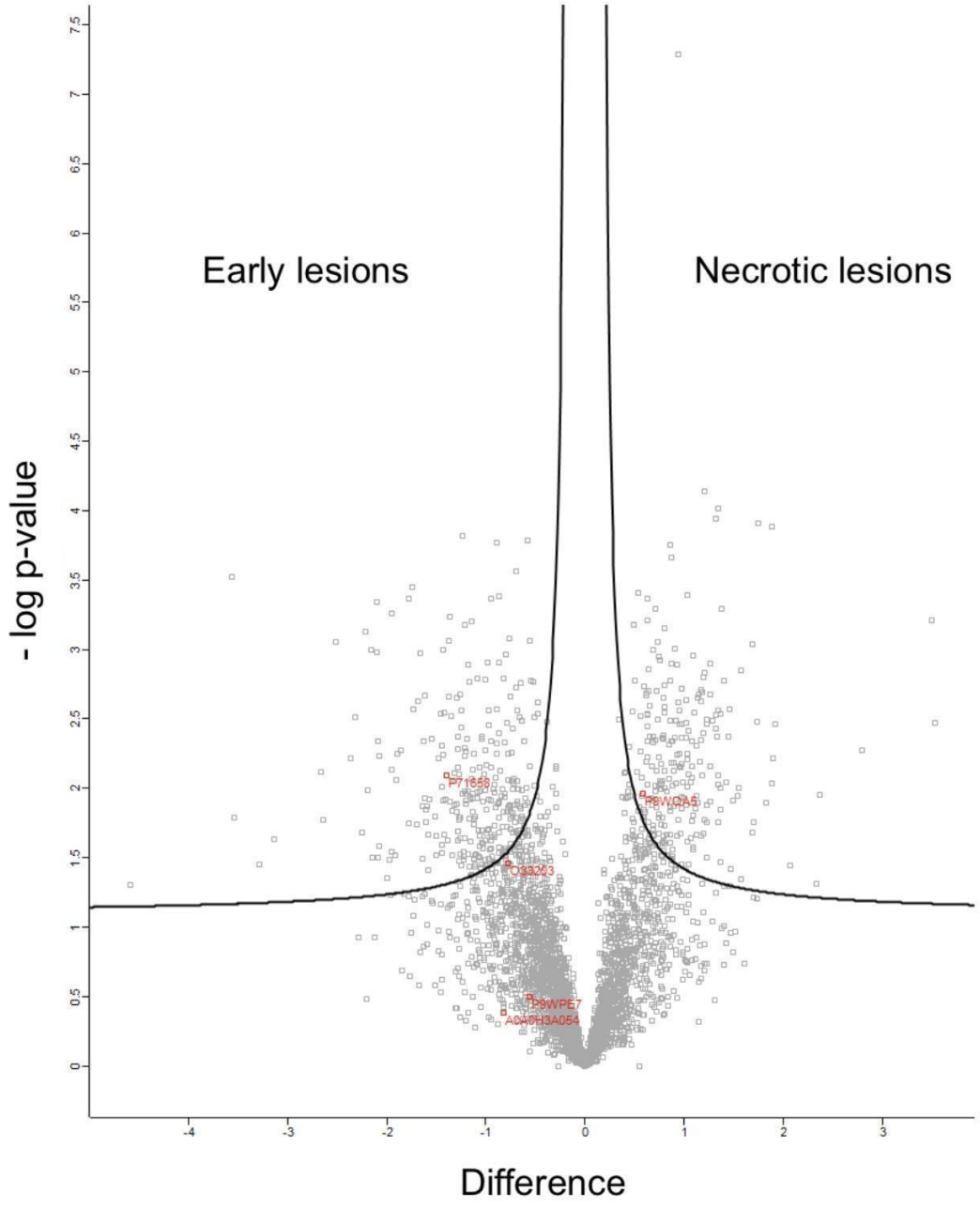


Figure 15: Volcano plot of all proteins found in both necrotic and early lesions. The x-axis shows the difference and the x-axis shows the -log p-value. The plot displays the necrotic lesion versus the early lesion with an FDR cut-off at a significance level of 1.2 and displays a clear separation of which proteins were found in the two lesions. The highlighted points in red are mycobacterial proteins. Those that are on the left are proteins that are mainly found in early

lesions and the ones on the right are mainly found in necrotic lesions. Graphical representation of the interaction was generated by Perseus software version 1.2.15.0.

In total there were 3641 proteins plotted on the diagram in Figure 15. There were a total of five proteins found and two were considered significant, P71658 in the early lesion and P9WQA5 in the necrotic lesion. Table 19 displays all five mycobacterial proteins with their accession key, gene name, bacteria and a short description of the protein. The two significant proteins are highlighted, and both belong to MTB.

Table 19: Table of all mycobacterial proteins found in both lesions including their description. Highlighted in yellow are the proteins that were found to be statistically significant when using the volcano plot, refer to Figure 15.

Accession	**Gene**	**Bacteria**	**Description**
P71658	mihF	*Mycobacterium tuberculosis* (strain ATCC 25618 / H37Rv)	Putative integration host factor mihF
P9WPE7	groEL2	*Mycobacterium tuberculosis* (strain ATCC 25618 / H37Rv)	60 kDa chaperonin 2; prevents aggregation of substrate proteins and promotes their refolding
A0A0H3A054	hup	*Mycobacterium avium* (strain 104)	DNA-binding protein HU homolog
P9WQA5	Rv2971	*Mycobacterium tuberculosis* (strain ATCC 25618 / H37Rv)	Uncharacterised oxidoreductase; belongs to the aldo/keto reductase family
O33203	cycA	*Mycobacterium tuberculosis* (strain ATCC 25618 / H37Rv)	Probable D-serine/alanine/glycine transporter protein cycA

3.5.1.1 mihF

Integration host factors are proteins that bind and bend to DNA with the role of transcriptional regulation and local DNA structural organisation in gram-negative bacteria [87]. However, mycobacterium tuberculosis is a gram positive bacteria [88]. Several recent research has been published that explored the role of protein mihF. One study found that mihF binds to DNA in a non-sequence specific manner, has the ability of bending and compacting the DNA and relaxes the strand [89]. A more recent study conducted in 2020 found similar results, where mihF, a rigid, alpha-helic protein that does not undergo structural rearrangement, binds only to double stranded DNA in solution with two binding sites and that, according to Atomic Force Microscopy, mihF unwinds and relaxes the DNA by introducing left-handed loops in supercoiled cosmids [90]. DNA is either supercoiled or relaxed and why supercoiled DNA is preferred due to compaction of the strand, allowing more DNA packaging within a cell, and its requirement for cell synthesis [91].

3.5.1.2 groEL2

Chaperonin proteins assist in folding, refolding, translocating and the regulation of their own genes [92]. MTB has two chaperonin genes; groEL1 and groEL2, however it has been found that only groEL2 is essential. Due to groEL2 being dimeric and lacking in ATPase activity, it does not function like normal chaperonin genes. Instead, groEL2 is antigenic and promotes cytokines interleukin-10 and tumour necrosis factor-alpha from monocytes from a CD14-independent manner [93]. It is able to detach itself from the bacteria, pass through a cell membrane and towards the mitochondria organelles. It blocks apoptosis by interacting with a member of the HSP 70 gene, therefore extending the life of the cell and increasing the longevity of MTB survival [94].

3.5.1.3 hup

HU proteins play an architectural role, are dimeric and involved in replication and compaction of DNA along with regulation of replication and recombination [95, 96]. HU proteins are nucleoid-associated proteins, which affect the transcriptional landscape of bacterial cells, and have a broad specificity when binding to DNA. A recent study conducted in 2017 found three

features that influence the specificity of the HU proteins binding to DNA; the salt-bridge formation creating conformational restrictions, increased conformational space of flexible side chains due to alanine enrichment and the nature of DNA binding residue which interacts differentially to DNA bases [97]. An under expression of the HU proteins results in decompaction of the nucleoid, whereas an overexpression leads to nucleoid condensation and DNA compaction, preventing nuclease digestion [95].

3.5.1.4 Rv2971

Aldo-keto reductases are a large superfamily, catalysing NADPH-dependent oxidoreductases and are involved in detoxifying a broad range of toxic metabolites [98]. According to several studies, MTB only has two putative aldo-keto reductases, Rv2971 and Rv2298, however it has been shown that only Rv2971 is essential for MTB growth [98-100]. One study suggests that Rv2971 may be involved in the detoxification of methylglyoxal, a highly toxic molecule produced in eukaryotes and prokaryotes, in MTB and that if there was an under expression of Rv2971, there would be a build up of methylglyoxal in the cell which would lead to terminal effects [99].

3.5.1.5 cycA

The protein cycA is a bacterial D-serine, L- and D-alanine, glycine transporter, induced by L-alanine with 12 helical trans-membrane domains from the amino acid transporter family [101, 102]. L- and D-alanine is necessary for MTB growth as it synthesises peptidoglycan in the mycobacterial cell wall, and an overexpression of D-alanine can increase resistant to the second line antibiotic for drug-resistant TB treatment, D-cycloserine [103]. Peptidoglycan assists in protection and preservation of the bacterial cell and is an important regulator of cell division, hence MTB growth [104, 105].

3.5.2 Protein abundance of the identified mycobacterial proteins

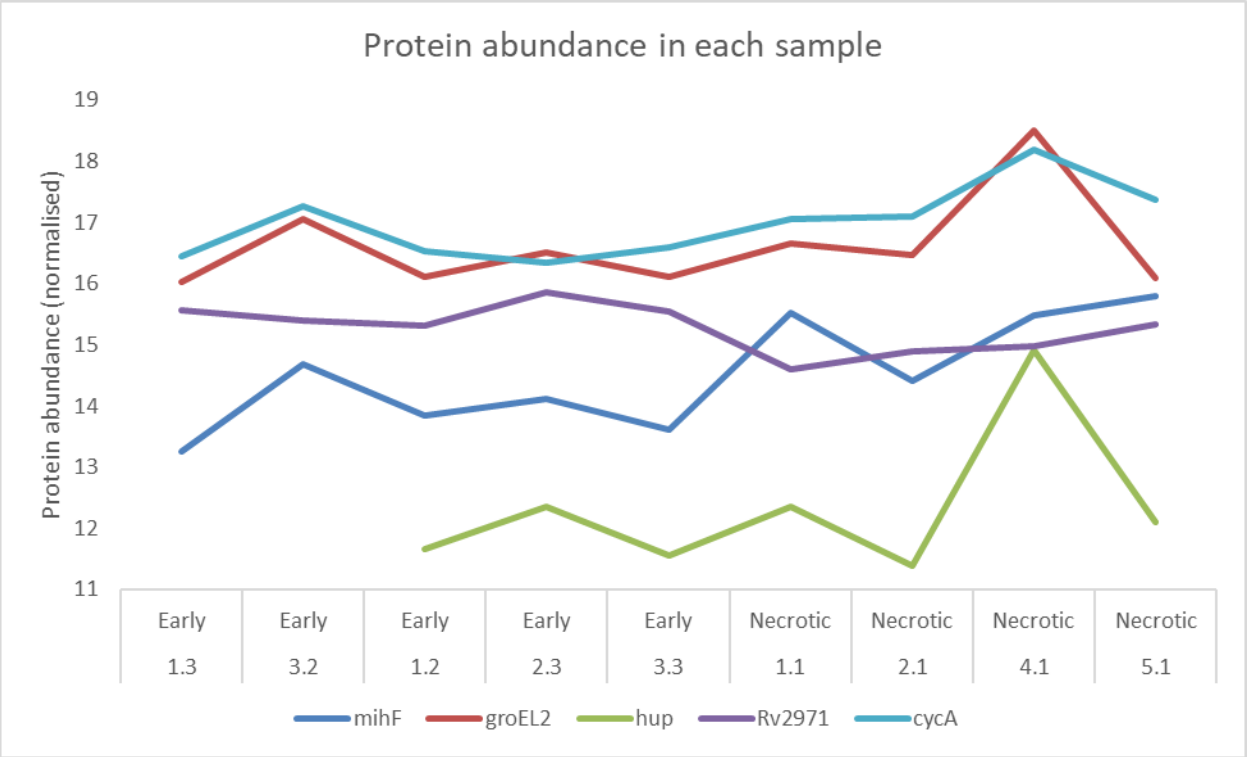


Figure 16: Line graph of the protein abundance in each sample. Y-axis displays protein abundance (normalised). X-axis displays lesions (early or necrotic), study number (1-5) and type of protein (mihF, groEL2, hup, Rv2971, cycA). Each line has a different colour corresponding to which type of protein it is. Generated using Excel version 2203.

There is a difference between the lesions, but within each lesion they are generally similar. The proteins mihF, groEL2, hup and cycA found more proteins in the samples from the necrotic lesions than the samples from the early lesions. On the other hand, all samples from the early lesions identified more Rv2917 than in samples from the necrotic lesions. The protein hup has the smallest abundance of proteins in all of the samples and was only identified in 60% of the early lesion samples.

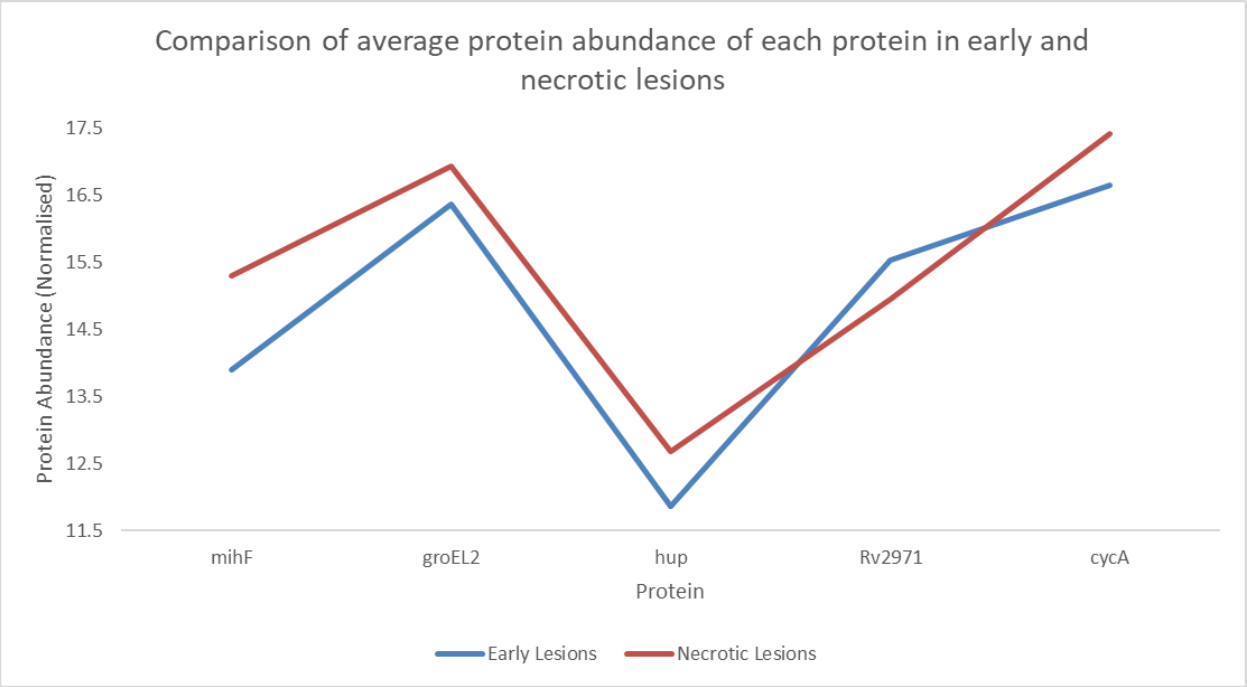


Figure 17: Line graph of a comparison of average protein abundance in the different proteins from early and necrotic lesions. Y-axis displays the protein abundance (normalised) and x-axis displays the study number. The blue line is the early lesions and the red line is necrotic lesions. Generated using Excel version 2203.

The early lesions have a lower amount of protein abundance compared to the necrotic lesions, except for the protein Rv2971, where the amount is higher in the early lesions compared to the necrotic lesions.

Table 20: Proteins and their protein abundance (normalised) averages in micrograms in all samples from early and necrotic lesions and the difference of necrotic lesions compared to the early lesions.

Gene	Early lesions	Necrotic lesions	Log2 fold change (necrotic - early)
mihF	13.901	15.302	1.401
groEL2	16.362	16.923	0.561

hup	11.859	12.684	0.825
Rv2971	15.530	14.949	- 0.581
cycA	16.639	17.417	0.778

The largest difference of protein abundance is with the protein mihF with a 1.401 difference. Protein groEL2 has the smallest difference with 0.561. All proteins have more protein in the necrotic lesions except for Rv2971.

3.5.3 Hierarchical clustering of the identified mycobacterial proteins

Along the left side is a list of proteins found in the sample and on the top are the different lesions with their case numbers. When looking at the colours, the more red means that there is a higher intensity and the more blue signifies a lower intensity. Due to the data not being imputed, the values that were missing have not been filled in and instead are in grey.

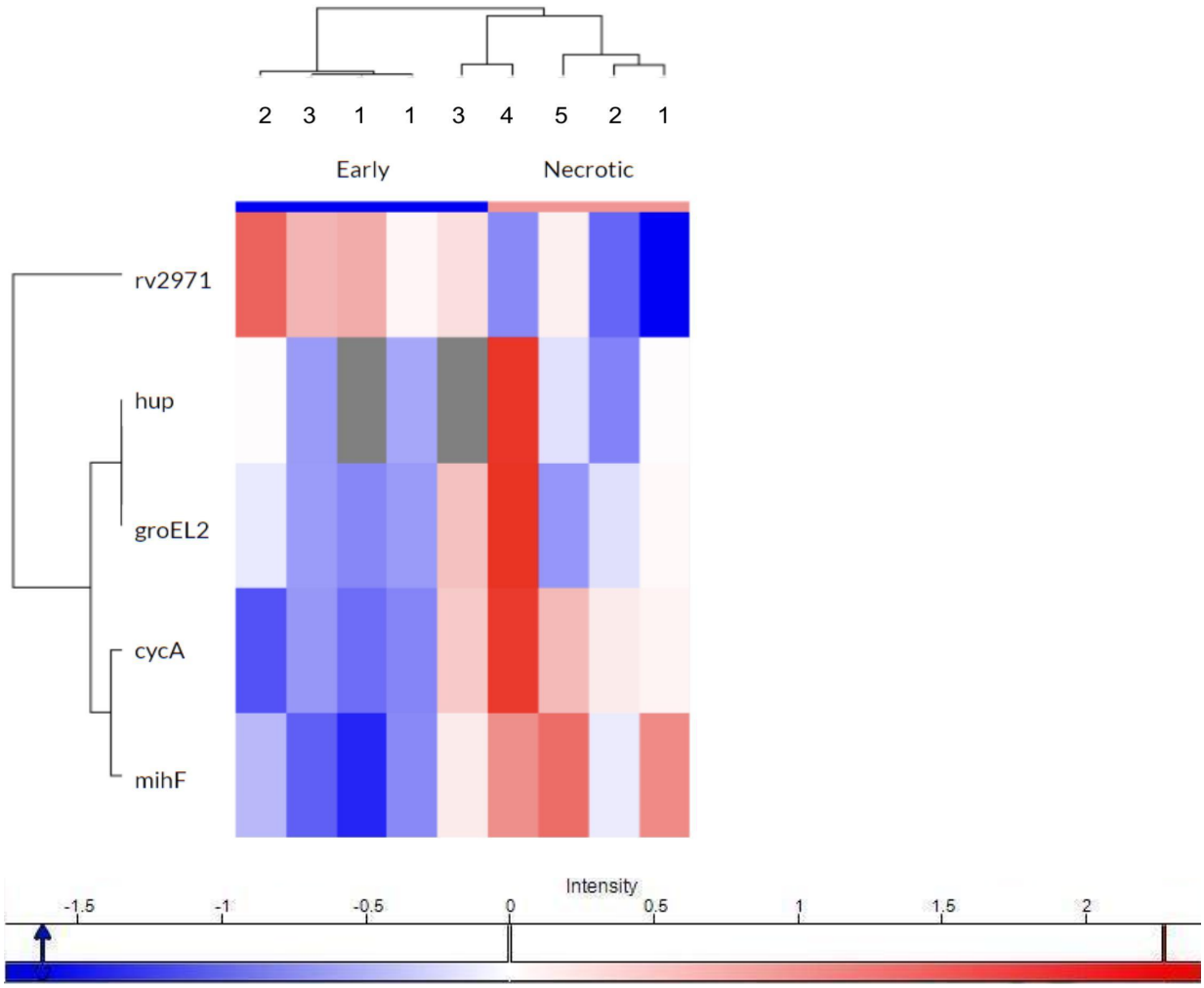


Figure 18: Supervised hierarchical clustering of mycobacterial proteins found in both early and necrotic lesions.

The samples have been grouped with their corresponding lesions in Figure 18, early lesions on the left and necrotic lesions on the right. There is higher intensity of Rv2971 in the early lesions compared to the necrotic lesions. The other four proteins; hup, groEL2, cycA and mihF, have a higher intensity in the necrotic lesions and a lower intensity in the early lesions.

Table 21: Table of all mycobacterium proteins found in both lesions, values were found using student t-test, early - necrotic lesions. (p-value)

Accession	Gene name	q-value combined	p-value	Difference	Test statistics
P71658	mihF	0.030	0.008	-1.4011	-3.6638
P9WPE7	groEL2	0.000	0.316	-0.5609	-1.0798
A0A0H3A054	hup	0.006	0.417	-0.8252	-0.8840
P9WQA5	Rv2971	0.001	0.011	0.5815	3.4356
O33203	cycA	0.007	0.034	-0.7781	-2.6211

As displayed in Table 21, the combined q-values are all less than 0.03 which implies that less than 3% of the data will result in false positives. The p-values for mihF, Rv2971 and cycA are shown to be significant as they are less than 0.05, meaning that the null hypothesis has been rejected and that there is less than 5% chance that the results were random. The t-test difference shows in values where the proteins interact more in. The values that are negative interact less in the early lesions and more in the necrotic lesions, whereas the values that are positive interact more in the early lesions and less in the necrotic lesions. T-test test statistics show how significant the differences between each group are, as shown, mihF and Rv2971 are significantly higher than the other proteins, making them significant.

3.5.4 Protein interaction network

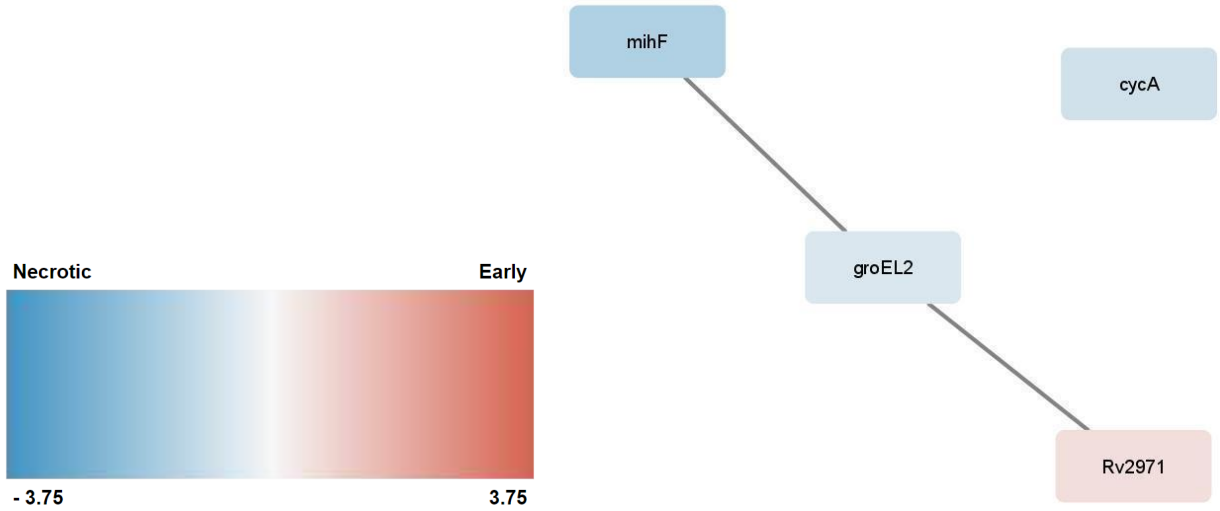


Figure 19: Network interaction analysis of the mycobacterium tuberculosis proteins found in all the samples with fold change. The more blue the nodes, the more differentially expressed it is in the necrotic lesion and the more red, the more differentially expressed it is in the early lesions. Graphical representation of the interaction was generated by Cytoscape software version 3.9.1.

Figure 19 shows the interactions found in the mycobacterium tuberculosis proteins with four proteins. Three of the proteins are more differentially expressed in the necrotic lesions, whereas the Rv2971 protein is more differentially expressed in the early lesions. GroEL2 and Rv2971 interact with one another. A closer look at the interaction between the mycobacterial proteins, mihF, groEL2 and Rv2971 have only been shown to interact due to being co-mentioned in PubMed articles.

4 Discussion

4.1 Results

4.1.1 Mycobacterial proteins

There was a clear difference between the early and necrotic lesions that was displayed with multiple graphs and backed up with statistical testing. Each table and figure displayed the same results with the five mycobacterial proteins, as two were continuously shown to be statistically significantly differentially expressed.

Five mycobacterial proteins were found in the lesions. Among these, protein RV2971 was significantly differentially expressed with more abundance in the early lesion compared to the necrotic lesions. This is involved in detoxification of methylglyoxal and methylglyoxal is involved with destruction of cells as it is highly toxic. Relative abundance of the Rv2971 protein implies that mycobacterial does not allow the host cell to get destroyed and hence the bacteria uses the host cells to evade immunity and go undetected. Patient 2 from this project is the oldest, at 45 years of age, and had been found to have a higher intensity of RV2971 in the early lesion and a significantly low intensity in the necrotic lesion. This further strengthens the possibility that RV2971 acts as a protector to the host as this case may have been infected by MTB a long time ago, whereas if we look at the youngest patient, 4, who had quite a low intensity of RV2971 in their necrotic sample, which could be due to the patient being an obvious case of disseminated disease. One study also theorised that the protein Rv2971 may also have a secondary role in resistance due to residues being found interacting with streptomycin, a first line anti-TB drug, far from the conserved site [106]. Earlier studies have shown that MTB can increase the expression of anti-apoptotic proteins, FasL and Bcl2, and it has been hypothesised that MTB may evade detection by the immune system by virtue of hiding in the macrophages which are unable to lyse the bacteria[107]. The histopathology of early lesions show that lesions are composed of vacuolated macrophages and there is paucity of lymphocytes in these lesions. Based on scientific literature, this protein, RV2971, could be associated with immune evasion of MTB. The early has been shown to regress in some cases, making those differentially expressed in early lesions of special interest to develop vaccines, which may lead to regression of early lesions and protection.

It is also interesting to note that the anti-TB drug Isoniazid inhibits this protein and so Rv2971 could also be a potential drug target [98].

The protein mihF was also significantly differentially expressed, however more abundant in the necrotic lesions as compared to the early lesions. This protein relaxes DNA strands, hence disrupting the function of DNA. Although, it is uncertain whether it binds to host or bacterial DNA. If this protein binds to the host, then it would interfere with the cellular process, perhaps leading to the destruction of the cells. Most likely, from the results in this project, mihF binds to the host as it is more differentially expressed in the necrotic lesions. So, this could be a candidate for host-directed therapies, where it could be possible to manipulate this protein to reduce tissue destruction during disease and could even improve and shorten treatment time.

The proteins groEL2, hup and cycA were not significantly differentially expressed, however are slightly more expressed in the necrotic lesions than the early lesions. groEL2 is a chaperone protein and blocks apoptosis by interacting with a member of the HSP 70 gene, therefore extending the life of the cell and increasing the longevity of MTB survival. A decreased expression of the hup protein results in decompaction of the nucleoid, whereas an overexpression leads to nucleoid condensation and DNA compaction, preventing nuclease digestion. The cycA protein synthesises peptidoglycan in the mycobacterial cell wall, which assists in protection and preservation of the bacterial cell and is an important regulator of cell division, hence MTB growth. An overexpression of cycA can increase resistance to the second line antibiotic for drug-resistant TB treatment, D-cycloserine [103]. These three play a role in assisting MTB survival and could potentially be used as biomarkers for TB, which are used to see the host's response to treatment of disease.

Patient 4 was five months old, presented cavities, had meningitis and her cause of death was due to TB, an obvious case of disseminated disease as the patient's immune system had not yet fully developed. This could explain why this case had an exaggerated expression of the hup, cycA and groEL2 proteins in the necrotic lesions as compared to the other four cases, as displayed in Figure 16 and Figure 18. However, there is the question of why only these three were high in this case and not all the proteins.

4.1.2 Bacterial proteins

Of the seven bacterial proteins that were significantly differentially expressed in the early lesions compared to the necrotic lesions, four of these were proteins from bacterias that were found only to early lesions. These proteins were lmo0282 from *Listeria monocytogenes senvar*, CRN1 from *Candida albicans*, HP_1024 from *Helicobacter pylori* and Blon_2093 from *Bifidobacterium longum subsp. infantis*.

Listeria monocytogenes senvar is involved with pulmonary infections in infants, pregnant women, immunocompromised or elderly individuals and are rarely found in adults [108-110]. All patients from this study were adults, except for patient 4, who was five months old. This case did not display a relatively high or low intensity in both lesions, which most likely means that lom0282 was consistently abundant in both lesions. This patient was also reported to have meningitis, and it is known that *Listeria monocytogenes* can cause tubercular meningitis with the help of MTB, which has been discussed to also be present in this patient [111].

Though the samples from this project were from the lungs, changes in the microbiota of the gut can highly influence the entire system of the host including the lungs and many researchers have been emphasising the importance of this in relation to TB[70, 112]. A study found that patients were 50% less likely to develop TB disease if the bacteria *Helicobacter pylori* was present in their gut flora [113]. A few other studies looked at *Helicobacter pylori* and its role with TB, finding similar results, that an abundance of *Helicobacter pylori* results in less likely to develop TB disease [71, 113]. This could explain the reason as to why there is a higher intensity of HP_1024 from *Helicobacter pylori* in the early lesions as compared to the necrotic lesions, as it acts as a protection against TB. Interestingly, patient 1 had a significantly high intensity of HP_1024 in one of their early lesion samples and was also reported to have not died from TB and instead from ulcer perforates, which has been established to be caused by *Helicobacter pylori* [114], further intensifying the theory that *Helicobacter pylori* may be protecting the host from TB, however does cause other detrimental effects on the body. Further research into this bacteria, particularly the HP_1024 protein could lead to finding a vaccine candidate.

Both *Candida albicans* and *Bifidobacterium infantis* are associated with allergies. A study using mice tested the significance of the gastrointestinal tract using *Candida albicans* and found that

allergic responses were induced [72]. The same study looked at *Bifidobacterium infantis* and found that this microbiota assisted in generating oral tolerance. Previous studies have shown the relationship between TB and allergies, but it is unclear whether TB patients are more susceptible to manifest allergies than healthy patients or if individuals with allergies are more susceptible to TB [115]. On the other hand, another study reported that those with allergies, specifically allergic rhinitis, are more protected from developing TB disease, possibly due to the TH2-mediated eosinophilic response [116]. TH2 releases cytokines that suppress macrophage functioning and instead focus on eosinophils, white blood cells that fight diseases and TB is known to suppress TH2, whereas allergies inhibit TH1 [117, 118]. Regardless of if allergies assist or fight TB disease from progressing, it is justifiable for these proteins from bacteria that are associated with allergies to be more present in the early lesions. Although, from the results found in this project, it would seem that it leans more towards allergies acting as something that may protect the host, because of the role of TH2, the protein CRN1 being an actin binding protein, which regulates architecture and structure of the cell [119], and the protein Blon_2093 from *Bifidobacterium infantis*, as it promotes health.

4.2 Methodology

This project aimed to establish a proof of principle for FFPE TB material from the pre-antibiotic era using proteomics and have found encouraging results where trends can easily be seen and explained between and within samples. However, there are a few things about the methodology that should be discussed.

4.2.1 Autopsy tissue material

Human tissue is the only relevant material to study the immunopathogenesis of the post-primary TB, as this form of disease occurs only in humans. None of the animal models in current use are relevant for the post-primary TB as all animals including non-human primates, produce only primary disease. However, there is scarcity of human TB material for research purposes. The availability of such material plunged after the introduction of antibiotics due to sharp decline in mortality. Most of the laboratories have not kept their old autopsy material systematically. This

has led researchers to rely on the animal models and has diverted focus from the pathogenesis of post-primary TB. Erroneously, the immunity in primary TB is considered relevant in all forms of TB leaving a large knowledge gap in the understanding of post-primary TB.

The Norwegian archival TB material is well-preserved and applicable for the advanced immunological and microbiological techniques. Accompanying autopsy, medical and epidemiological records make this material a unique resource, as such material is not available in other countries. It is a unique opportunity to be able to investigate relevant human material using modern techniques. It will potentially give new and invaluable insight into the true immunopathology of the post-primary TB and would be a leap forward towards understanding the distinct immune processes of post primary TB.

4.2.2 Laser microdissection

Laser microdissection made it possible to accurately separate lesions from the same sample in order to conduct proteomics and therefore provide results that were able to be comparable. It is also quicker than other micro-dissecting methods and preserves tissue morphology. One requirement to use a laser microdissection is being able to identify cells of interest based on morphological characteristics, discussion and learning this skill was necessary in order to continue with the use of the laser microdissection. One limitation was that it was difficult to properly identify and distinguish the sections under the machine's microscope and camera. To reduce this error, a HE-staining of the case and a photo of the stained slide was accessible to help visualise and identify which lesions to cut, an example of this was in an earlier slide. There is also the risk of under quality tissues due to dehydration and to prevent this, liquid coverslips were applied onto the slides after staining and a PEN membrane glass slide was used. Another limitation of laser microdissection is that it is highly time consuming and an expensive tool.[64, 120-122]

4.2.3 Database search

The database search was created through researching published papers that suggest bacteria involved with the lung. However, there is the question of what other bacteria could have been

added, as it is uncertain as to what would be found from these samples and there is a high chance that there are bacteria not known to be involved in the lung. It is important to acknowledge that the database used for this project was quite basic and should be improved.

4.2.4 Proteomics

As for Proteomics, drawing the line of what is deemed important and what is perhaps not is quite difficult. The use of statistics and knowledge of physiological processes helps guide researchers in this case. Proteomics is quite complex and determining exactly which procedures to follow is determined based on the type of tissues worked with.

Since the tissue used in this experiment is formalin fixed an extensive protocol for lysis is needed. Formalin in an aqueous environment binds to amines, purines, and thiols forming methyl derivatives that are covalently bound to tissue [123]. This results in crosslinking of proteins, RNA and DNA leading to fixation of the tissue. In order to break the crosslinks extensive boiling and sonication is needed.

A step that was added to the proteomics methodology was the TMT labelling, allowing the running of all the samples through the mass spectrometer at the same time, which reduces biases, as each run on the mass spectrometer will be slightly different, for example contamination.

Off line fractionation, a process which splits the samples, could have been added as this would create more separation between the samples, giving cleaner results and improve identification and quantification of the proteins. The reason this was not done for this project was due to the small amount of sample taken from each case and the risk of losing the sample was too high.

4.2.5 Certainty of results and normalisation

The combination of the statistics displayed on Table 16 and Figures 13 and 14 show that the confidence these proteins have been correctly identified is high. Normalisation was first done in Thermo Proteome Discoverer; the abundances were normalised to the same total peptide amount per channel and scaled, so that the average abundance per protein and peptide totalled to 100. Samples were again normalised in Perseus.

4.2.6 emPAI

The emPAI scores that have been found for this project are significant, as it further strengthens the existence of four of the mycobacterial proteins. emPAI only looks at the concentration of the protein and disregards the other molecules and chemicals mixed into the sample during the proteomics process. It is not surprising when compared to other research findings that the numbers they find are higher, as they have grown and isolated the bacteria through culture, therefore their samples would be purer, whereas the samples used in this project came from human FFPE blocks from the pre-antibiotic era. On the other hand, when the emPAI from other studies is found to be lower, this could be due to the specific protein and their function. Perhaps, the bacteria need a human host in order to express certain proteins. As for the protein, cycA, it is not certain why there were no articles that looked at this protein's emPAI or why there are very few articles about this specific protein in general.[78]

4.2.7 Comparison with other studies

There are a few studies that used a comparable methodology to this project. A recent study conducted in 2020 that used FFPE blocks to compare TB and *Mycobacterium avium* complex lung disease found mycobacterium proteins in caseous necrotic regions, a unique form of cell death with a cheese-like appearance, in the granulomas [73]. Another similar study from 2017 analysed MTB lineage 7 and 4, detecting less abundance of proteins associated with slow growth and virulence in lineage 7 as compared to lineage 4 [124]. A review published in 2018 discusses the use of proteomics in studying MTB, emphasising that there is potential in this methodology and that more study is needed in this area as the medical advancement possibilities are crucial and possible through proteomics [125].

4.3 Limitations and biases of the study and mitigation

It is important that bias and limitations are discussed and considered when analysing the results. There is a natural selection bias, as only people dying in the hospitals were autopsied and included in the study. This is also a cross-sectional study where we cannot study the temporal relationship of lesions, however, progression of lesions has been described in old literature when

scientists had more possibility to study and correlate lesions seen on serial x-rays with pathology seen on autopsies.

There were a total of five patients used in this project. The sample size was small and heterogeneous, hence results having low statistical power and potential biases and therefore it is difficult to generalise the findings. Studies must be done on larger and more homogenous samples to make any conclusive statements about the importance of differentially expressed proteins found. However, these samples are quite promising as we can easily group the lesions together and have seen relationships and trends with the proteins that have been found.

As parts of this project were conducted in a lab, it is inevitable that human error will occur during the process. To prevent this as best as possible, practice samples were used before using the real samples in order to understand and learn from experienced practitioners and supervision was also accessible.

4.4 Further and future studies

As this project was meant for a proof of principle and only used five cases with a total of eleven samples, further studies could add many more samples to strengthen results. A future study could possibly also look at other types of lesions such as fibroblast lesions, which is the process of necrosis healing, and normal lung tissues. Comparing these could also give a more in-depth insight into the different proteins being expressed at different stages.

There have been many articles that have discussed and emphasised the importance of the gut microbiota and how it interacts and affects the other organs, in particular the lungs. Further studies could delve into this with proteomics and compare the relationship between the two in TB infected samples.

With technology advancing every day, one that has recently been talked about is nanopore. It is a small device that is more direct in the analysis of proteins, that needs less sample preparation, has a potential for long sequence reads and is relatively inexpensive, therefore ideal to place in lower economically developed countries [126-129]. The mass spectrometer has many limitations and with this new technology it could improve the study of proteins.

4.5 Global impact

This methodology has been shown to work with FFPE blocks from the pre-antibiotic era. Further studies could establish significant proteins that could potentially be used to create protein biomarkers for detection and monitoring of TB, something that is rather difficult in the world today. It could also be used to produce possible vaccine targets or better treatments, essential especially in lower economically developed countries who have a high burden of TB. To study TB and find candidates for vaccines and treatments, early lesions are what needs to be focused on as there is clearly something that triggers TB to progress into a disease.

5 Conclusion

The results obtained from this proteomics study identified five mycobacterial proteins and of these, two were significantly differentially expressed; RV2971 in the early lesions and mihF in the necrotic lesions. Overall, based on the results that have been accumulated from just five cases, this pilot study has established a proof of principle that is feasible to detect and identify very low quantities of mycobacterial proteins present in the abundant host proteins, which are otherwise not detectable by routine methods. This methodology shows promise in strengthening theories and bringing forth new knowledge that could be used to build biomarkers, develop new vaccines and improve treatment for TB, advancing medicine for TB across the globe.

Author contributions

Tehmina Mustafa was the main supervisor of this project, she designed and revised this project. Syeda Mariam Riaz and Even Birkeland were co-supervising this project and both assisted and revised the project. Syeda Marium Riaz taught staining and the epidemiology of the tissues. Even Birkeland taught proteomics and the analysis of the project.

Ethics

All data is anonymised. The project has been approved by the Regional Ethical Committee REK, ref 2019/45.

Funding

This project is a part of a larger project and used funding from it and applied for extra funding and support from the University of Bergen. Partial funding of the project was provided by the Blakstad Maarschalk og Helbings legat for bekjempelse av tuberkulose og kreft.

Conflict of interest

The author declares no conflict of interest.

References

1. *Tuberculosis*. ERS Monograph. 2018: European Respiratory Society.
2. Pezzella, A.T., *History of Pulmonary Tuberculosis.* Thorac Surg Clin, 2019. **29**(1): p. 1-17.
3. *Global Tuberculosis Report 2020: Executive Summary*. 2020, World Health Organization: Geneva.
4. *Core Curriculum on Tuberculosis: What the Clinicians Should Know*. 2013, Centers for Disease Control and Prevention.
5. *Total number of people who developed any form of TB (including DR-TB and TB-HIV co-infection)*. 2020, Stop TB Partnership: Geneva.
6. Organization, W.H., *Incidence, New and relapse TB cases notified, HIV-positive TB incidence*. 2022: Geneva.
7. *Total number of people who died from any form of TB (including DR-TB and TB-HIV co-infection)*. 2020, Stop TB Partnership: Geneva.
8. Organization, W.H., *HIV-negative TB mortality*. 2022, World Health Organization: Geneva.
9. *Transforming our World: The 2030 Agenda for Sustainable Development*. 2015, United Nations: Geneva.
10. Pai, M., et al., *Tuberculosis.* Nat Rev Dis Primers, 2016. **2**: p. 16076.
11. Marais, B.J. and H.S. Schaaf, *Tuberculosis in children.* Cold Spring Harb Perspect Med, 2014. **4**(9): p. a017855.

12. Houda Ben Ayed, M.K., Chakib Marrakchi, Khaoula Rekik, Fatma Hammami, Fatma Smaoui, Mariem Ben Hmida, Sourour Yaich, Imed Maaloul, Jamel Damak, Mounir Ben Jemaa, *Extrapulmonary Tuberculosis: Update on the Epidemiology, Risk Factors and Prevention Strategies.* International Journal of Tropical Diseases, 2018. **1**(1).
13. Padmanesan Narasimhan, J.W., Chandini Raina MacIntyre, Dilip Mathai, *Risk Factors for Tuberculosis.* Pulmonary Medicine, 2013.
14. Campbell, I.A. and O. Bah-Sow, *Pulmonary tuberculosis: diagnosis and treatment.* BMJ, 2006. **332**(7551): p. 1194-7.
15. Bhalla, A.S., et al., *Chest tuberculosis: Radiological review and imaging recommendations.* Indian J Radiol Imaging, 2015. **25**(3): p. 213-25.
16. Assael, R., J. Cervantes, and G. Barrera, *Smears and cultures for diagnosis of pulmonary tuberculosis in an asymptomatic immigrant population.* Int J Gen Med, 2013. **6**: p. 777-9.
17. Asghar, M.U., et al., *Sputum Smear and Culture-negative Tuberculosis with Associated Pleural Effusion: A Diagnostic Challenge.* Cureus, 2018. **10**(10): p. e3513.
18. Kivihya-Ndugga, L., et al., *Comparison of PCR with the routine procedure for diagnosis of tuberculosis in a population with high prevalences of tuberculosis and human immunodeficiency virus.* J Clin Microbiol, 2004. **42**(3): p. 1012-5.
19. Menon, P.K., K. Kapila, and V.C. Ohri, *Polymerase Chain Reaction and Advances in Infectious Disease Diagnosis.* Med J Armed Forces India, 1999. **55**(3): p. 229-231.
20. Salian, N.V., et al., *Polymerase chain reaction to detect Mycobacterium tuberculosis in histologic specimens.* Am J Respir Crit Care Med, 1998. **158**(4): p. 1150-5.
21. Amin, I., et al., *PCR could be a method of choice for identification of both pulmonary and extra-pulmonary tuberculosis.* BMC Res Notes, 2011. **4**: p. 332.
22. Richeldi, L., *An update on the diagnosis of tuberculosis infection.* Am J Respir Crit Care Med, 2006. **174**(7): p. 736-42.
23. Sharma, S.K., et al., *Comparison of TST and IGRA in Diagnosis of Latent Tuberculosis Infection in a High TB-Burden Setting.* PLoS One, 2017. **12**(1): p. e0169539.
24. *Guidelines on the management of latent tuberculosis infection*. 2015, World Health Organization: Geneva.

25. Piccazzo, R., F. Paparo, and G. Garlaschi, *Diagnostic accuracy of chest radiography for the diagnosis of tuberculosis (TB) and its role in the detection of latent TB infection: a systematic review.* J Rheumatol Suppl, 2014. **91**: p. 32-40.
26. Luca, S. and T. Mihaescu, *History of BCG Vaccine.* Maedica (Bucur), 2013. **8**(1): p. 53-8.
27. Payam Nahid, S.E.D., Narges Alipanah, Pennan M. Barry, Jan L. Brozek, Adithya Cattamanchi, Lelia H. Chaisson, Richard E. Chaisson, Charles L. Daley, Malgosia Grzemska, Julie M. Higashi, Christine S. Ho, Philip C. Hopewell, Salmaan A. Keshavjee, Christian Lienhardt, Richard Menzies, Cynthia Merrifield, Masahiro Narita, Rick O'Brien, Charles A. Peloquin, Ann Raftery, Jussi Saukkonen, H. Simon Schaaf, Giovanni Sotgiu, Jeffrey R. Starke, Giovanni Battista Migliori, Andrew Vernon, *Executive Summary: Official American Thoracic Society/Centers for Disease Control and Prevention/Infectious Diseases Society of America Clinical Practice Guidelines: Treatment of Drug-Susceptible Tuberculosis.* Clinical Infectious Diseases, 2016. **63**(7): p. 853-867.
28. Horsburgh, C.R., Jr., C.E. Barry, 3rd, and C. Lange, *Treatment of Tuberculosis.* N Engl J Med, 2015. **373**(22): p. 2149-60.
29. Wei, L., et al., *Host-directed therapy, an untapped opportunity for antimalarial intervention.* Cell Rep Med, 2021. **2**(10): p. 100423.
30. Zumla, A., et al., *Host-directed therapies for infectious diseases: current status, recent progress, and future prospects.* Lancet Infect Dis, 2016. **16**(4): p. e47-63.
31. Kaufmann, S.H.E., et al., *Host-directed therapies for bacterial and viral infections.* Nat Rev Drug Discov, 2018. **17**(1): p. 35-56.
32. Yong, Y.K., et al., *Immune Biomarkers for Diagnosis and Treatment Monitoring of Tuberculosis: Current Developments and Future Prospects.* Front Microbiol, 2019. **10**: p. 2789.
33. Goletti, D., et al., *Tuberculosis Biomarkers: From Diagnosis to Protection.* Infect Dis Rep, 2016. **8**(2): p. 6568.
34. Munro, S.A., et al., *Patient adherence to tuberculosis treatment: a systematic review of qualitative research.* PLoS Med, 2007. **4**(7): p. e238.

35. Alipanah, N., et al., *Adherence interventions and outcomes of tuberculosis treatment: A systematic review and meta-analysis of trials and observational studies.* PLoS Med, 2018. **15**(7): p. e1002595.
36. *Implementing the WHO Stop TB Strategy: A Handbook for National Tuberculosis Control Programmes*. 2008.
37. Jamila Rashid, O.O.T., Indu Ahluwalia, Stella Chungong *Disparities in Infectious Diseases among Women in Developing Countries*. 2004. **10**, DOI: 10.3201/eid1011.040624_12.
38. Worley, H. *Chronic Diseases Beleaguer Developing Countries*. 2006.
39. *Incidence of tuberculosis (per 100,000) people) - Norway*, W.H. Organization, Editor. 2020: The World Bank.
40. Twyman, R.M., *Proteomics.* Encyclopedia of Applied Ethics, 2012: p. 642-649.
41. Kleiner, M., *Metaproteomics: Much More than Measuring Gene Expression in Microbial Communities.* mSystems, 2019. **4**(3).
42. Aslam, B., et al., *Proteomics: Technologies and Their Applications.* J Chromatogr Sci, 2017. **55**(2): p. 182-196.
43. Elena Burillo, I.J., Diego Martínez-López, Emilio Camafeita, Jesus Vazquez, Jose L. Martin–Ventura, *Chapter 11 - The Application of Proteomic Techniques in the Study of HDL Particle Characterization and Biomarker Discovery.* The HDL Handbook, 2017: p. 231-255.
44. Pran Kishore Deb, O.A.-A., Balakumar Chandrasekaran, Anant Paradkar, Rakesh K. Tekade, *Chapter 16 - Protein/Peptide Drug Delivery Systems: Practical Considerations in Pharmaceutical Product Development.* Basic Fundamentals of Drug Delivery, 2019: p. 651-684.
45. Allen, G., *Laboratory Techniques in Biochemistry and Molecular Biology*, ed. R.H.B. T.S. Work. Vol. 9. 1981, Beckenham: Elsevier/Norh-Holland Biomedical Press.
46. Shemetov, A.A., I. Nabiev, and A. Sukhanova, *Molecular interaction of proteins and peptides with nanoparticles.* ACS Nano, 2012. **6**(6): p. 4585-602.

47. Mary Joseph, H.M.T., Ashim K. Mitra, *Chapter 7 - Peptide and Protein-Based Therapeutic Agents.* Emerging Nanotechnologies for Diagnostics, Drug Delivery and Medical Devices, 2017: p. 145-167.
48. Glish, G.L. and R.W. Vachet, *The basics of mass spectrometry in the twenty-first century.* Nat Rev Drug Discov, 2003. **2**(2): p. 140-50.
49. Domon, B. and R. Aebersold, *Mass spectrometry and protein analysis.* Science, 2006. **312**(5771): p. 212-7.
50. Hossain, M., *The Mass Spectrometer and Its Components.* Selected Reaction Monitoring Mass Spectrometry (SRM-MS) in Proteomics, 2020: p. 17-52.
51. Zhang, L. and J.E. Elias, *Relative Protein Quantification Using Tandem Mass Tag Mass Spectrometry.* Methods Mol Biol, 2017. **1550**: p. 185-198.
52. Dayon, L. and J.C. Sanchez, *Relative protein quantification by MS/MS using the tandem mass tag technology.* Methods Mol Biol, 2012. **893**: p. 115-27.
53. Navarrete-Perea, J., et al., *Streamlined Tandem Mass Tag (SL-TMT) Protocol: An Efficient Strategy for Quantitative (Phospho)proteome Profiling Using Tandem Mass Tag-Synchronous Precursor Selection-MS3.* J Proteome Res, 2018. **17**(6): p. 2226-2236.
54. Hung, C.W. and A. Tholey, *Tandem mass tag protein labeling for top-down identification and quantification.* Anal Chem, 2012. **84**(1): p. 161-70.
55. Steen, H. and M. Mann, *The ABC's (and XYZ's) of peptide sequencing.* Nat Rev Mol Cell Biol, 2004. **5**(9): p. 699-711.
56. Cleveland, J.P. and J.R. Rose, *Identification of b-/y-ions in MS/MS spectra using a two stage neural network.* Proteome Sci, 2013. **11**(Suppl 1): p. S4.
57. Renard, B.Y., et al., *When less can yield more - Computational preprocessing of MS/MS spectra for peptide identification.* Proteomics, 2009. **9**(21): p. 4978-84.
58. Litlekalsoy, J., et al., *Time-dependent biological differences in molecular markers of high-grade urothelial cancer over 7 decades (ras proteins, pTEN, uPAR, PAI-1 and MMP-9).* Virchows Arch, 2012. **461**(5): p. 541-51.
59. Bertelsen, B.I., et al., *HPV subtypes in cervical cancer biopsies between 1930 and 2004: detection using general primer pair PCR and sequencing.* Virchows Arch, 2006. **449**(2): p. 141-7.

60. Malte Bayer, L.A., Christoph Schliemann, Wolfgang Hartmann, Simone König, *Are formalin-fixed and paraffin-embedded tissues fit for proteomic analysis?* Journal of Mass Spectrometry, 2019. **55**(8).
61. Teemu, R., *"Tuberculosis-threatened children": the rise and fall of a medical concept in Norway, c.1900-1960.* Med Hist, 2008. **52**(3): p. 347-64.
62. Kummari, E., S.X. Guo-Ross, and J.B. Eells, *Laser capture microdissection--a demonstration of the isolation of individual dopamine neurons and the entire ventral tegmental area.* J Vis Exp, 2015(96): p. e52336.
63. Espina, V., et al., *Laser-capture microdissection.* Nat Protoc, 2006. **1**(2): p. 586-603.
64. Chung, S.H. and W. Shen, *Laser capture microdissection: from its principle to applications in research on neurodegeneration.* Neural Regen Res, 2015. **10**(6): p. 897-8.
65. Hunter, R.L., *The Pathogenesis of Tuberculosis-The Koch Phenomenon Reinstated.* Pathogens, 2020. **9**(10).
66. Hunter, R.L., *Tuberculosis as a three-act play: A new paradigm for the pathogenesis of pulmonary tuberculosis.* Tuberculosis (Edinb), 2016. **97**: p. 8-17.
67. Orme, I.M. and R.J. Basaraba, *The formation of the granuloma in tuberculosis infection.* Semin Immunol, 2014. **26**(6): p. 601-9.
68. Buczak, K., et al., *Spatially resolved analysis of FFPE tissue proteomes by quantitative mass spectrometry.* Nat Protoc, 2020. **15**(9): p. 2956-2979.
69. O'Dwyer, D.N., R.P. Dickson, and B.B. Moore, *The Lung Microbiome, Immunity, and the Pathogenesis of Chronic Lung Disease.* J Immunol, 2016. **196**(12): p. 4839-47.
70. Namasivayam, S., et al., *The Microbiome and Tuberculosis: Early Evidence for Cross Talk.* mBio, 2018. **9**(5).
71. Charissa C Naidoo, G.R.N., Benjamin G Wu, Gerhard Walzl, Robin M Warren, Leopoldo N Segal, Grant Theron, *The microbiome and tuberculosis: state of the art, potential applications, and defining the clinical research agenda.* Respiratory Medicine, 2019. **7**(10): p. 892-906.
72. James M. Beck, V.B.Y., Gary B. Huffnagle, *The microbiome of the lung.* Translational Research, 2012. **160**(4): p. 258-266.

73. Seto, S., et al., *Proteomic Profiling Reveals the Architecture of Granulomatous Lesions Caused by Tuberculosis and Mycobacterium avium Complex Lung Disease.* Front Microbiol, 2019. **10**: p. 3081.
74. Pahal, P., V. Rajasurya, and S. Sharma, *Typical Bacterial Pneumonia*, in *StatPearls*. 2022: Treasure Island (FL).
75. Consortium, T.U., *UniProt: the universal protein knowledgebase in 2021.* Nucleic Acids Research, 2020. **49**(D1): p. D480–D489.
76. Orsburn, B.C., *Proteome Discoverer-A Community Enhanced Data Processing Suite for Protein Informatics.* Proteomes, 2021. **9**(1).
77. Tyanova, S., et al., *The Perseus computational platform for comprehensive analysis of (prote)omics data.* Nat Methods, 2016. **13**(9): p. 731-40.
78. Ishihama, Y., et al., *Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein.* Mol Cell Proteomics, 2005. **4**(9): p. 1265-72.
79. Szklarczyk, D., et al., *STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets.* Nucleic Acids Res, 2019. **47**(D1): p. D607-D613.
80. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome Res, 2003. **13**(11): p. 2498-504.
81. Chin, J.H., *Tuberculous meningitis: Diagnostic and therapeutic challenges.* Neurol Clin Pract, 2014. **4**(3): p. 199-205.
82. de Souza, G.A., et al., *Using a label-free proteomics method to identify differentially abundant proteins in closely related hypo- and hypervirulent clinical Mycobacterium tuberculosis Beijing isolates.* Mol Cell Proteomics, 2010. **9**(11): p. 2414-23.
83. de Souza, G.A., et al., *Bacterial proteins with cleaved or uncleaved signal peptides of the general secretory pathway.* J Proteomics, 2011. **75**(2): p. 502-10.
84. Tucci, P., et al., *Integrative proteomic and glycoproteomic profiling of Mycobacterium tuberculosis culture filtrate.* PLoS One, 2020. **15**(3): p. e0221837.
85. Borsuk, S., et al., *Identification of proteins from tuberculin purified protein derivative (PPD) by LC-MS/MS.* Tuberculosis (Edinb), 2009. **89**(6): p. 423-30.

86. Thomas, L., et al., *Metabolic switches and adaptations deduced from the proteomes of Streptomyces coelicolor wild type and phoP mutant grown in batch culture.* Mol Cell Proteomics, 2012. **11**(2): p. M111 013797.
87. Mangan, M.W., et al., *The integration host factor (IHF) integrates stationary-phase and virulence gene expression in Salmonella enterica serovar Typhimurium.* Mol Microbiol, 2006. **59**(6): p. 1831-47.
88. Fu, L.M. and C.S. Fu-Liu, *Is Mycobacterium tuberculosis a closer relative to Gram-positive or Gram-negative bacterial pathogens?* Tuberculosis (Edinb), 2002. **82**(2-3): p. 85-90.
89. Mishra, A., et al., *Integration host factor of Mycobacterium tuberculosis, mIHF, compacts DNA by a bending mechanism.* PLoS One, 2013. **8**(7): p. e69985.
90. Odermatt, N.T., et al., *Structural and DNA binding properties of mycobacterial integration host factor mIHF.* J Struct Biol, 2020. **209**(3): p. 107434.
91. William R. Bauer, F.H.C.C., James H. White, *Supercoiled DNA.* Scientific American, 1980. **243**: p. 118-133.
92. A.J.L. Macario, E.C.d.M., *Chaperone Proteins and Chaperonopathies.* Encyclopedia of Stress, 2007: p. 438-444.
93. Stapleton, M.R., et al., *Mycobacterium tuberculosis WhiB1 represses transcription of the essential chaperonin GroEL2.* Tuberculosis (Edinb), 2012. **92**(4): p. 328-32.
94. Joseph, S., et al., *Mycobacterium tuberculosis Cpn60.2 (GroEL2) blocks macrophage apoptosis via interaction with mitochondrial mortalin.* Biol Open, 2017. **6**(4): p. 481-488.
95. Grove, A., *Functional Evolution of Bacterial Histone-Like HU Proteins.* Current Issues in molecular biology, 2011. **13**: p. 1-12.
96. Tilly, K., et al., *Isolation of Borrelia burgdorferi genes encoding homologues of DNA-binding protein HU and ribosomal protein S20.* Microbiology (Reading), 1996. **142 (Pt 9)**: p. 2471-9.
97. Dey, D., V. Nagaraja, and S. Ramakumar, *Structural and evolutionary analyses reveal determinants of DNA binding specificities of nucleoid-associated proteins HU and IHF.* Mol Phylogenet Evol, 2017. **107**: p. 356-366.

98. Shahine, A., et al., *A structural characterization of the isoniazid Mycobacterium tuberculosis drug target, Rv2971, in its unliganded form.* Acta Crystallogr F Struct Biol Commun, 2014. **70**(Pt 5): p. 572-7.
99. Scoble, J., et al., *Crystal structure and comparative functional analyses of a Mycobacterium aldo-keto reductase.* J Mol Biol, 2010. **398**(1): p. 26-39.
100. Sassetti, C.M., D.H. Boyd, and E.J. Rubin, *Genes required for mycobacterial growth defined by high density mutagenesis.* Mol Microbiol, 2003. **48**(1): p. 77-84.
101. Hong, W., L. Chen, and J. Xie, *Molecular basis underlying Mycobacterium tuberculosis D-cycloserine resistance. Is there a role for ubiquinone and menaquinone metabolic pathways?* Expert Opin Ther Targets, 2014. **18**(6): p. 691-701.
102. Chen, J.M., et al., *A point mutation in cycA partially contributes to the D-cycloserine resistance trait of Mycobacterium bovis BCG vaccine strains.* PLoS One, 2012. **7**(8): p. e43467.
103. Chen, J., et al., *Identification of novel mutations associated with cycloserine resistance in Mycobacterium tuberculosis.* J Antimicrob Chemother, 2017. **72**(12): p. 3272-3276.
104. Vollmer, W., *Chapter 6 - Peptidoglycan.* Molecular Medical Microbiology, 2015. **1**: p. 105-124.
105. Y.-P. Hsu, X.M., M.S. Van Nieuwenhze, *Chapter 1 - Methods for visualization of peptidoglycan biosynthesis.* Methods in Microbiology, 2016. **43**: p. 3-48.
106. Sharma, P., et al., *Proteomic analysis of streptomycin resistant and sensitive clinical isolates of Mycobacterium tuberculosis.* Proteome Sci, 2010. **8**: p. 59.
107. S. J. Mogga, T.M., L. Sviland, R. Nilsen, *Increased Bcl-2 and reduced Bax expression in infected macrophages in slowly progressive primary murine Mycobacterium tuberculosis infection.* Scandinavian Journal of Immunology, 2002: p. 383-391.
108. Ananthraman, A., R.H. Israel, and C.R. Magnussen, *Pleural-pulmonary aspects of Listeria monocytogenes infection.* Respiration, 1983. **44**(2): p. 153-7.
109. Theocharis Koufakis, M.C., Anastasios Margaritis, Maria Tsiakalou, Ioannis Gabranis, *Pneumonia by Listeria monocytogenes: A Common Infection by an Uncommon Pathogen.* Case Reports in Infectious Diseases, 2015.

110. Morgand, M., et al., *Listeria monocytogenes-associated respiratory infections: a study of 38 consecutive cases.* Clin Microbiol Infect, 2018. **24**(12): p. 1339 e1-1339 e5.
111. Paciorek, M., et al., *Differences and similarities in clinical manifestations of Listeria monocytogenes and Mycobacterium tuberculous meningitis.* Przegl Epidemiol, 2020. **74**(2): p. 326-335.
112. Sencio, V., M.G. Machado, and F. Trottein, *The lung-gut axis during viral respiratory infections: the impact of gut dysbiosis on secondary disease outcomes.* Mucosal Immunol, 2021. **14**(2): p. 296-304.
113. Perry, S., et al., *Infection with Helicobacter pylori is associated with protection against tuberculosis.* PLoS One, 2010. **5**(1): p. e8804.
114. Gisbert, J.P. and J.M. Pajares, *Helicobacter pylori infection and perforated peptic ulcer prevalence of the infection and role of antimicrobial treatment.* Helicobacter, 2003. **8**(3): p. 159-67.
115. Ellertsen, L.K., et al., *Allergic sensitisation in tuberculosis patients at the time of diagnosis and following chemotherapy.* BMC Infect Dis, 2009. **9**: p. 100.
116. Yii, A.C., et al., *Asthma, Sinonasal Disease, and the Risk of Active Tuberculosis.* J Allergy Clin Immunol Pract, 2019. **7**(2): p. 641-648 e1.
117. Mayumi Saeki, T.N., Noriko Kitamura, Takachika Hiroi, Akio Mori, Osamu Kaminuma *Potential Mechanisms of T Cell-Mediated and Eosinophil-Independent Bronchial Hyperresponsiveness.* Allergic Airway Disease: Molecular Immunology, Pathogenesis and Inflammation, 2019.
118. Sabbagh, D.K., R. Beasley, and G.B. Marks, *The Immunological Mysteries of Tuberculosis.* J Allergy Clin Immunol Pract, 2019. **7**(2): p. 649-650.
119. dos Remedios, C.G., et al., *Actin binding proteins: regulation of cytoskeletal microfilaments.* Physiol Rev, 2003. **83**(2): p. 433-73.
120. Fend, F. and M. Raffeld, *Laser capture microdissection in pathology.* J Clin Pathol, 2000. **53**(9): p. 666-72.
121. Jensen, E., *Laser-capture microdissection.* Anat Rec (Hoboken), 2013. **296**(11): p. 1683-7.

122. Mahalingam, M., *Laser Capture Microdissection: Insights into Methods and Applications.* Methods Mol Biol, 2018. **1723**: p. 1-17.
123. Thavarajah, R., et al., *Chemical and physical basics of routine formaldehyde fixation.* J Oral Maxillofac Pathol, 2012. **16**(3): p. 400-5.
124. Yimer, S.A., et al., *Comparative Proteomic Analysis of Mycobacterium tuberculosis Lineage 7 and Lineage 4 Strains Reveals Differentially Abundant Proteins Linked to Slow Growth and Virulence.* Front Microbiol, 2017. **8**: p. 795.
125. Bisht, D., et al., *Recent insights into Mycobacterium tuberculosis through proteomics and implications for the clinic.* Expert Rev Proteomics, 2019. **16**(5): p. 443-456.
126. Branton, D., et al., *The potential and challenges of nanopore sequencing.* Nat Biotechnol, 2008. **26**(10): p. 1146-53.
127. Eisenstein, M., *Oxford Nanopore announcement sets sequencing sector abuzz.* Nat Biotechnol, 2012. **30**(4): p. 295-6.
128. Venkatesan, B.M. and R. Bashir, *Nanopore sensors for nucleic acid analysis.* Nat Nanotechnol, 2011. **6**(10): p. 615-24.
129. Deamer, D., M. Akeson, and D. Branton, *Three decades of nanopore sequencing.* Nat Biotechnol, 2016. **34**(5): p. 518-24.

Appendix

Appendix 1: Instrument method HPLC

Instrument Method

D:\01_Eclipse-DATA\SERVICE\Beatrice_Normann\02212022_Eclipse_P_xxx\Raw\Beatrice_F2.raw

---- Overview ----

Name: New Instrument Method

Comment:

Run time: 140.000 [min]

Instrument: Eclipsenano on thermo-4fpicdhd

Description:

---- Script ----

initial Instrument Setup

PumpModule.LoadingPump.%A.Equate: "%A"

PumpModule.LoadingPump.%B.Equate: "%B"

PumpModule.LoadingPump.%C.Equate: "%C"

PumpModule.LoadingPump.Pressure.LowerLimit: 0 [bar]

PumpModule.LoadingPump.Pressure.UpperLimit: 400 [bar]

PumpModule.LoadingPump.MaximumFlowRampUp: 10 [µl/min²]

PumpModule.LoadingPump.MaximumFlowRampDown: 10 [µl/min²]

PumpModule.NC_Pump.%A.Equate: "%A"

PumpModule.NC_Pump.%B.Equate: "%B"

PumpModule.NC_Pump.Pressure.LowerLimit: 0 [bar]

PumpModule.NC_Pump.Pressure.UpperLimit: 700 [bar]

PumpModule.NC_Pump.MaximumFlowRampUp: 0.0500 [µl/min²]

PumpModule.NC_Pump.MaximumFlowRampDown: 0.300 [µl/min²]

ColumnOven.TempCtrl: On

ColumnOven.Temperature.Nominal: 40.0 [°C]

ColumnOven.Temperature.LowerLimit: 20.0 [°C]

ColumnOven.Temperature.UpperLimit: 75.0 [°C]

ColumnOven.EquilibrationTime: 0.5 [min]

ColumnOven.ReadyTempDelta: 1.0 [°C]

Sampler.LowDispersionMode: Off

Sampler.WashSpeed: 4.000 [µl/s]

Sampler.WashVolume: 50.000 [µl]

Sampler.PunctureDepth: 7.000 [mm]

Sampler.SampleHeight: 1.000 [mm]

Sampler.WasteSpeed: 4.000 [µl/s]

Sampler.DispenseDelay: 2.000 [s]

Sampler.DispSpeed: 2.000 [µl/s]

Sampler.DrawSpeed: 0.050 [µl/s]

Sampler.DrawDelay: 5.000 [s]

Sampler.RinseBetweenReinjections: No

Sampler.FlushVolume: 5.000 [µl]

Sampler.TransVialPunctureDepth: 8.000 [mm]

Sampler.TransLiquidHeight: 4.000 [mm]

Sampler.TransportVialCapacity: Unlimited

Sampler.LastTransportVial: G1

Sampler.FirstTransportVial: G1

Sampler.InjectMode: ulPickUp

Sampler.LoopWashFactor: 2.000

Sampler.PumpDevice: "LoadingPump"

Sampler.TempCtrl: On

Sampler.Temperature.Nominal: 10.0 [°C]

Sampler.ReadyTempDelta: 2.0 [°C]

Sampler.Temperature.LowerLimit: 4.0 [°C]

Sampler.Temperature.UpperLimit: 20.0 [°C]

ColumnOven.ValveRight: 6_1

0.000 [min] Equilibration

PumpModule.LoadingPump.Flow.Nominal: 5.000 [µl/min]

PumpModule.LoadingPump.%B.Value: 0.0 [%]

PumpModule.LoadingPump.%C.Value: 0.0 [%]

PumpModule.LoadingPump.Curve: 5

PumpModule.NC_Pump.Flow.Nominal: 0.250 [µl/min]

PumpModule.NC_Pump.%B.Value: 5.0 [%]

PumpModule.NC_Pump.Curve: 5

0.000 [min] Inject Preparation

Wait PumpModule.LoadingPump.Ready And PumpModule.NC_Pump.Ready And ColumnOven.Ready And Sampler.Ready

0.000 [min] Inject

Sampler.Inject

0.000 [min] Start Run

PumpModule.NC_Pump.NC_Pump_Pressure.AcqOn

PumpModule.LoadingPump.LoadingPump_Pressure.AcqOn

0.000 [min] Run

PumpModule.LoadingPump.Flow.Nominal: 5.000 [µl/min]

PumpModule.LoadingPump.%B.Value: 0.0 [%]

PumpModule.LoadingPump.%C.Value: 0.0 [%]

PumpModule.LoadingPump.Curve: 5

PumpModule.NC_Pump.Flow.Nominal: 0.250 [µl/min]

PumpModule.NC_Pump.%B.Value: 5.0 [%]

PumpModule.NC_Pump.Curve: 5

5.000 [min]

PumpModule.NC_Pump.Flow.Nominal: 0.250 [µl/min]

PumpModule.NC_Pump.%B.Value: 5.0 [%]

PumpModule.NC_Pump.Curve: 5

ColumnOven.ValveRight: 1_2

6.000 [min]

PumpModule.NC_Pump.Flow.Nominal: 0.250 [µl/min]

PumpModule.NC_Pump.%B.Value: 7.0 [%]

PumpModule.NC_Pump.Curve: 5

10.000 [min]

Sampler.Wash

90.000 [min]

PumpModule.NC_Pump.Flow.Nominal: 0.250 [µl/min]

PumpModule.NC_Pump.%B.Value: 22.0 [%]

PumpModule.NC_Pump.Curve: 5

100.000 [min]

PumpModule.NC_Pump.Flow.Nominal: 0.250 [µl/min]

PumpModule.NC_Pump.%B.Value: 28.0 [%]

PumpModule.NC_Pump.Curve: 5

105.000 [min]

PumpModule.NC_Pump.Flow.Nominal: 0.250 [µl/min]

PumpModule.NC_Pump.%B.Value: 80.0 [%]

PumpModule.NC_Pump.Curve: 5

120.000 [min]

PumpModule.NC_Pump.Flow.Nominal: 0.250 [µl/min]

PumpModule.NC_Pump.%B.Value: 80.0 [%]

PumpModule.NC_Pump.Curve: 5

125.000 [min]

PumpModule.NC_Pump.Flow.Nominal: 0.250 [µl/min]

PumpModule.NC_Pump.%B.Value: 5.0 [%]

PumpModule.NC_Pump.Curve: 5

ColumnOven.ValveRight: 6_1

140.000 [min] Stop Run

PumpModule.NC_Pump.NC_Pump_Pressure.AcqOff

PumpModule.LoadingPump.LoadingPump_Pressure.AcqOff

Appendix 2: Processing Method for Thermo Proteome Discoverer

Result name: Beatrice_F

Result file: D:\Even\Beatrice_Normann\Tuberculosis first search22022022\Beatrice_F-(2).msf

Description: -

Workflow based on template: processing_wf_tmt16plex_IRS_dyn_modif

Creation date: 3/2/2022 10:10:02 AM

Created with Discoverer version: 2.5.0.400

The workflow tree:

|-(0) Spectrum Files RC

|-(1) Spectrum Selector

|-(2) Precursor Detector

|-(3) Sequest HT

|-(4) Percolator

|-(5) Reporter Ions Quantifier

Processing node 0: Spectrum Files RC

---

1. Search Settings:

- File Name(s) (Hidden):

D:\Even\Beatrice_Normann\02212022_Eclipse_P_xxx\Raw\Beatrice_F1.raw

D:\Even\Beatrice_Normann\02212022_Eclipse_P_xxx\Raw\Beatrice_F1_20220222171619.raw

D:\Even\Beatrice_Normann\02212022_Eclipse_P_xxx\Raw\Beatrice_F2.raw

D:\Even\Beatrice_Normann\02212022_Eclipse_P_xxx\Raw\Beatrice_F2_20220222194346.raw

D:\Even\Beatrice_Normann\02212022_Eclipse_P_xxx\Raw\Beatrice_F3.raw

D:\Even\Beatrice_Normann\02212022_Eclipse_P_xxx\Raw\Beatrice_F3_20220222221120.raw

D:\Even\Beatrice_Normann\02212022_Eclipse_P_xxx\Raw\Beatrice_F4.raw

D:\Even\Beatrice_Normann\02212022_Eclipse_P_xxx\Raw\Beatrice_F4_20220223003853.raw

- Protein Database: All databases.fasta; Contaminants.fasta

- Enzyme Name: Trypsin (Full)

- Precursor Mass Tolerance: 20 ppm

- Fragment Mass Tolerance: 0.5 Da

- Static Peptide N-Terminus: TMTpro / +304.207 Da (Any N-Terminus)

- 1. Static Modification: Carbamidomethyl / +57.021 Da (C)

- 2. Static Modification: TMTpro / +304.207 Da (K)

2. Regression Settings:

- Regression Model: Non-linear Regression

- Parameter Tuning: Coarse

Processing node 1: Spectrum Selector

---

1. General Settings:

- Precursor Selection: Use MS1 Precursor

- Use Isotope Pattern in Precursor Reevaluation: True

- Provide Profile Spectra: Automatic

2. Spectrum Properties Filter:

- Lower RT Limit: 0

- Upper RT Limit: 0

- First Scan: 0

- Last Scan: 0

- Lowest Charge State: 0

- Highest Charge State: 0

- Min. Precursor Mass: 350 Da

- Max. Precursor Mass: 6000 Da

- Total Intensity Threshold: 0

- Minimum Peak Count: 1

3. Scan Event Filters:

- Mass Analyzer: Is FTMS

- MS Order: Is MS2

- Activation Type: Is HCD

- Min. Collision Energy: 0

- Max. Collision Energy: 1000

- Scan Type: Is Full

- Polarity Mode: Is +

4. Peak Filters:

- S/N Threshold (FT-only): 1.5

5. Replacements for Unrecognized Properties:

- Unrecognized Charge Replacements: Automatic

- Unrecognized Mass Analyzer Replacements: FTMS

- Unrecognized MS Order Replacements: MS2

- Unrecognized Activation Type Replacements: HCD

- Unrecognized Polarity Replacements: +

- Unrecognized MS Resolution@200 Replacements: 60000

- Unrecognized MSn Resolution@200 Replacements: 30000

6. Precursor Pattern Extraction:

- Precursor Clipping Range Before: 2.5 Da

- Precursor Clipping Range After: 5.5 Da

Processing node 2: Precursor Detector

---

1. General:

- S/N Threshold: 1.5

- S/N Threshold for most abundant peak: 1.5

Processing node 3: Sequest HT

---

1. Input Data:

- Protein Database: All databases.fasta; Contaminants.fasta

- Enzyme Name: Trypsin (Full)

- Max. Missed Cleavage Sites: 2

- Min. Peptide Length: 6

- Max. Peptide Length: 150

- Max. Number of Peptides Reported: 10

2. Tolerances:

- Precursor Mass Tolerance: 10 ppm

- Fragment Mass Tolerance: 0.02 Da

- Use Average Precursor Mass: False

- Use Average Fragment Mass: False

3. Spectrum Matching:

- Use Neutral Loss a Ions: True

- Use Neutral Loss b Ions: True

- Use Neutral Loss y Ions: True

- Use Flanking Ions: True

- Weight of a Ions: 0

- Weight of b Ions: 1

- Weight of c Ions: 0

- Weight of x Ions: 0

- Weight of y Ions: 1

- Weight of z Ions: 0

4. Dynamic Modifications:

- Max. Equal Modifications Per Peptide: 3

- Max. Dynamic Modifications Per Peptide: 4

- 1. Dynamic Modification: Oxidation / +15.995 Da (M)

- 2. Dynamic Modification: TMTpro / +304.207 Da (K)

5. Dynamic Modifications (peptide terminus):

- 1. N-Terminal Modification: TMTpro / +304.207 Da (N-Terminus)

7. Static Modifications:

- 2. Static Modification: Carbamidomethyl / +57.021 Da (C)

Processing node 4: Percolator

---

1. Target/Decoy Strategy:

- Target/Decoy Selection: Concatenated

- Validation based on: q-Value

2. Input Data:

- Maximum Delta Cn: 0.05

- Maximum Rank: 0

3. FDR Targets:

- Target FDR (Strict): 0.01

- Target FDR (Relaxed): 0.05

Processing node 5: Reporter Ions Quantifier

---

1. Peak Integration:

- Integration Tolerance: 20 ppm

- Integration Method: Most Confident Centroid

2. Scan Event Filters:

- Mass Analyzer: FTMS

- MS Order: MS2

- Activation Type: HCD

- Min. Collision Energy: 0

- Max. Collision Energy: 1000

Appendix 3: Consensus Method for Thermo Proteome Discoverer

Result name: Beatrice_F

Result file: D:\Even\Beatrice_Normann\Tuberculosis first search22022022\Beatrice_F-(2).pdResult

Description: Result filtered for high confident peptides, with enhanced peptide and protein annotations. Add FASTA file with common contaminants to the Protein Marker node. Quan abundances are normalized to the same total peptide amount per channel and scaled, so that the average abunce per protein and peptide is 100.

Workflow based on template: CWF_Comprehensive_Enhanced Annotation_Reporter_Quan

Creation date: 3/2/2022 10:10:05 AM

Created with Discoverer version: 2.5.0.400

The workflow tree:

|-(0) MSF Files

|-(1) PSM Grouper

|-(2) Peptide Validator

|-(3) Peptide and Protein Filter

|-(4) Protein Scorer

|-(7) Protein FDR Validator

|-(5) Protein Grouping

|-(6) Peptide in Protein Annotation

|-(9) Protein Marker

|-(10) Reporter Ions Quantifier

Post-processing nodes:

|-(11) Result Statistics

|-(12) Display Settings

|-(13) Data Distributions

Processing node 0: MSF Files

---

1. Storage Settings:

- Spectra to Store: Identified or Quantified

- Feature Traces to Store: All

2. Merging of Identified Peptide and Proteins:

- Merge Mode: Globally by Search Engine Type

3. FASTA Title Line Display:

- Reported FASTA Title Lines: Best match

- Title Line Rule: standard

4. PSM Filters:

- Maximum Delta Cn: 0.05

- Maximum Rank: 0

- Maximum Delta Mass: 0 ppm

Hidden Parameters:

- MSF File(s): D:\Even\Beatrice_Normann\Tuberculosis first search22022022\Beatrice_F-(2).msf

---

Processing node 1: PSM Grouper

---

1. Peptide Group Modifications:

- Site Probability Threshold: 75

Processing node 2: Peptide Validator

---

1. General Validation Settings:

- Validation Mode: Automatic (Control peptide level error rate if possible)

- Target FDR (Strict) for PSMs: 0.01

- Target FDR (Relaxed) for PSMs: 0.05

- Target FDR (Strict) for Peptides: 0.01

- Target FDR (Relaxed) for Peptides: 0.05

2. Specific Validation Settings:

- Validation Based on: q-Value

- Target/Decoy Selection for PSM Level FDR Calculation Based on Score: Automatic

- Reset Confidences for Nodes without Decoy Search (Fixed score thresholds): False

---

Processing node 3: Peptide and Protein Filter

---

1. Peptide Filters:

- Peptide Confidence At Least: High

- Keep Lower Confident PSMs: False

- Minimum Peptide Length: 6

- Remove Peptides Without Protein Reference: False

2. Protein Filters:

- Minimum Number of Peptide Sequences: 1

- Count Only Rank 1 Peptides: False

- Count Peptides Only for Top Scored Protein: False

Processing node 4: Protein Scorer

No parameters

Processing node 7: Protein FDR Validator

---

1. Confidence Thresholds:

- Target FDR (Strict): 0.01

- Target FDR (Relaxed): 0.05

Processing node 5: Protein Grouping

---

1. Protein Grouping:

- Apply strict parsimony principle: True

Processing node 6: Peptide in Protein Annotation

---

1. Flanking Residues:

- Annotate Flanking Residues of the Peptide: True

- Number Flanking Residues in Connection Tables: 1

2. Modifications in Peptide:

- Protein Modifications Reported: Only for Master Proteins

3. Modifications in Protein:

- Modification Sites Reported: All And Specific

- Minimum PSM Confidence: High

- Report Only PTMs: True

4. Positions in Protein:

- Protein Positions for Peptides: Only for Master Proteins

Processing node 9: Protein Marker

---

1. Contaminant Database:

- Protein Database: Contaminants.fasta

2. Additional Marker Database:

- Protein Database: 020982_7252uniprot-mycobacterium+tuberculosis-filtered-reviewed yes.fasta

5. Annotate Species:

- As Species Map: False

- As Species Names: False

6. Mark Additional Entities:

- Annotation Groups: False

- Pathway Groups: False

- Modification Sites: True

- Peptide Isoform Groups: True

Processing node 10: Reporter Ions Quantifier

---

1. General Quantification Settings:

- Peptides to Use: Unique + Razor

- Consider Protein Groups for Peptide Uniqueness: True

- Use Shared Quan Results: True

- Reject Quan Results with Missing Channels: False

2. Reporter Quantification:

- Reporter Abundance Based On: Intensity

- Apply Quan Value Corrections: True

- Co-Isolation Threshold: 50

- Average Reporter S/N Threshold: 10

- SPS Mass Matches [%] Threshold: 65

- Minimum Channel Occupancy [%] Threshold: 0

3. Normalization and Scaling:

- Normalization Mode: Total Peptide Amount

- Scaling Mode: None

4. Exclude Peptides from Protein Quantification:

- For Normalization: Use All Peptides

- For Protein Roll-Up: Use All Peptides

- For Pairwise Ratios: Exclude Modified

5. Quan Rollup and Hypothesis Testing:

- Protein Ratio Calculation: Protein Abundance Based

- Maximum Allowed Fold Change: 100

- Imputation Mode: None

- Hypothesis Test: ANOVA (Individual Proteins)

6. Quan Ratio Distributions:

- 1st Fold Change Threshold: 2

- 2nd Fold Change Threshold: 4

- 3rd Fold Change Threshold: 6

- 4th Fold Change Threshold: 8

- 5th Fold Change Threshold: 10

Processing node 11: Result Statistics

No parameters

Processing node 12: Display Settings

---

1. General:

- Filter Set:

    ### Filter Set MasterProteinFilter.filterset contains the following filters:

    ### Row Filter for TargetProtein:

Master is equal to Master

###

'magellan filter set' 1 'MasterProteinFilter.filterset' FiltersetProperties 1 'LastFileName' 'C:\Users\frank.berg\Desktop\MasterProteinFilter.filterset' Filter 'TargetProtein' 1 NARY_AND 1 = FilterConditionProperties 1 'NamedComparableFilterCondition/DisplayPropertyHint' 'Master' property 'Thermo.PD.EntityDataFramework.MasterProteinAssessment, Thermo.Magellan.EntityDataFramework' 'IsMasterProtein' constant 'Thermo.PD.EntityDataFramework.MasterProteinAssessment, Thermo.Magellan.EntityDataFramework' 'IsMasterProtein'

Processing node 13: Data Distributions

---

1. ID Distributions (Bottom-up):

- Peptides to Use: Only unique peptides based on protein groups