BOSTON UNIVERSITY
GRADUATE SCHOOL OF ARTS AND SCIENCES
AND
COLLEGE OF ENGINEERING

Dissertation

RECONSTRUCTION OF LIPID METABOLISM REGULATORY NETWORK IN
MYCOBACTERIUM TUBERCULOSIS

by

IRINA GLOTOVA
B.S., Lomonosov Moscow State University, 2009

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
2014
Approved by

First Reader

James E. Galagan, PhD
Associate Professor of Biomedical Engineering and Microbiology

Second Reader

Daniel Segre, PhD
Associate Professor of Biomedical Engineering

Third Reader

Igor Kramnik, PhD
Associate Professor of Medicine
RECONSTRUCTION OF LIPID METABOLISM REGULATORY NETWORK IN
MYCOBACTERIUM TUBERCULOSIS

(Order No. )

IRINA GLOTOVA

Boston University Graduate School of Arts and Sciences
and College of Engineering, 2014

Major Professor: James E. Galagan, Assistant Professor of Bioengineering

ABSTRACT

Lipid metabolism plays a prominent role in the survival of *Mycobacterium tuberculosis* (MTB) in both the macrophage and mammalian hosts. A central question in the reconstruction of lipid metabolism in MTB is the key regulatory programs responsible for changes in this metabolism under different conditions. One of the most studied conditions is hypoxia (oxygen deprivation) as adaptations to hypoxia are thought to play an important role in MTB pathogenesis and latency. To identify temporal trends during a hypoxic time course and associate them with possible regulators, expression data was clustered into paths using DREM algorithm and available ChIP-Seq data for over 80 transcription factors (TFs). The degree to which expression patterns might reflect the direct action of transcription factors was assessed by evaluating the consistency between the path, the expression of each TF binding genes in the path, and the predicted regulatory role of the TF from the regulatory network. DosR was correctly identified as an activator of the path corresponding almost entirely to its regulon. Rv0081 was found to be a candidate high level regulator broadly predictive of the overall expression of sets of genes during hypoxia and re-aeration.
An essential part of the continued survival of MTB in the host is due to its adaptation to the phagosomal compartment of the macrophage where the bacterium faces hypoxic as well as other stresses. Using the same approach and several time course expression data sets available for MTB in macrophage cultures, macrophage temporal expression models were built and compared to the hypoxia model. Similar trends were found in the expression of genes involved in respiration, cholesterol catabolism and methylcitrate cycle. In contrast, a group of genes responsible for the synthesis of complex cell wall lipids (PAT/DAT and SL-1) were up-regulated in the macrophage models and downregulated in the hypoxic model. PhoP was predicted as a potential main regulator of these genes as a result of pH change which takes place in the macrophage environment but not during hypoxia.
# Table of Contents

## Chapter 1. Introduction

1.1. *Mycobacterium tuberculosis* (MTB) – the causative agent for tuberculosis (TB)...
1

1.2. Latency, dormancy and persistence ................................................................. 1

1.3. Modeling persistence in animals ........................................................................ 2

1.4. Macrophage infection ......................................................................................... 3

1.4.1. Formation of the granuloma .......................................................................... 3

1.4.2. Mycobacterial prevention of phagosome-lysosome fusion ............................ 4

1.4.3. Systems for regulation of MTB metabolic and growth state in the granuloma 8

1.4.4. MTB defense mechanisms against oxidative and nitrosative stress ............. 9

1.4.5. Difference between murine and human macrophages .................................. 11

1.5. Modeling persistence *in vitro* ........................................................................... 12

1.6. Signature genes involved in hypoxic response .................................................. 13

1.7. Transcriptional regulation in latent phase ......................................................... 14

1.7.1. DosR ........................................................................................................... 14

1.7.2. KstR .......................................................................................................... 15

1.7.3. PhoP .......................................................................................................... 16

1.7.4. WhiB3 ....................................................................................................... 18

1.7.5. Sigma factors ............................................................................................ 19

1.8. Lipid metabolism .............................................................................................. 23
1.8.1. Cholesterol metabolism .................................................................23
1.8.2. Fatty acid metabolism .................................................................24
1.8.3. Mycolic acid metabolism .............................................................25
1.8.4. TAG accumulation and catabolism .............................................27
1.8.5. Complex lipid metabolism (SL-1, PAT/DAT, PDIM) .....................28

1.9. Project overview ..............................................................................33

Chapter 2. Identification of regulators and regulatory programs underlying gene
expression during hypoxia and re-aeration .............................................34

2.1. Introduction .....................................................................................34

2.2. Prediction of regulatory networks ....................................................34

2.2.1. General network inference methods .............................................34

2.2.2. Network prediction methods for time series data sets ..................37

2.2.3. DREM (Dynamic Regulatory Events Miner) ..............................45

2.3. Data sets ..........................................................................................50

2.4. Data processing and statistical methods .........................................50

2.4.1. Microarray data processing .........................................................50

2.4.2. TF binding list .............................................................................51

2.4.3. DREM parameters .......................................................................52

2.5. Regulatory programs predicted during hypoxia .............................52

2.6. Validation of key regulators and regulatory programs predicted by DREM ....55
2.7. Identification of Rv0081 as a candidate high-level regulator broadly predictive of the overall expression of a set of genes during hypoxia and re-aeration ..................58

Chapter 3. Comparison of regulatory programs during hypoxia and macrophage infection ........................................................................................................................................61

3.1. Introduction ........................................................................................................................................................................................................61

3.2. Profiling of MTB under conditions other than hypoxia.................................................................61

3.2.1. Transcriptional response in macrophages .......................................................................................61

3.2.2. Transcriptional response to reactive nitrogen and oxygen species ..................................65

3.2.3. Transcriptional response to nutrient starvation ........................................................................65

3.2.4. Transcriptional response to acidic conditions ...........................................................................66

3.2. Data sets............................................................................................................................................................................................................66

3.3. Data processing and statistical methods .........................................................................................67

3.3.1. Microarray data processing ......................................................................................................67

3.3.2. Comparison between expression data sets .................................................................................67

3.4. Pathway enrichment of hypoxia and macrophage models obtained with DREM..68

3.5. Regulators predicted to be responsible for the expression of genes in the early stages of both hypoxia and macrophage models .........................................................................................................................69

3.6. Genes, involved in the synthesis of SL-1, PAT/DAT and regulated by PhoP, have opposite expression patterns in hypoxia and macrophage models..........................................................70

3.7. Additional regulators whose action can explain the specific behavior of genes responsible for SL-1, PAT/DAT synthesis in hypoxia versus macrophage infection ...72
Chapter 4. Discussion..................................................................................................................74

4.1. Conclusions .........................................................................................................................74

4.2. Future directions ..................................................................................................................75

Appendices..................................................................................................................................76

Bibliography..................................................................................................................................79

Curriculum Vitae..........................................................................................................................94
List of Tables

Table 1. Number of replicates per experiment in the hypoxia/re-aeration expression data set.................................................................51

Table 2. DREM TF predictions for hypoxia and macrophage time courses (Appendix A)..............................................................................76
List of Figures

Figure 1. Mycobacterial cell wall envelope depicting some of the essential methyl-branched lipids
Figure 2. Schema of the DREM method
Figure 3. Hypoxia profiling
Figure 4. DREM paths during hypoxia/re-aeration
Figure 5. TFs predicted by DREM to be responsible for separation in gene expression patterns during the hypoxia time course
Figure 6. DREM path 1-3 regulated by DosR
Figure 7. DREM paths regulated by KstR
Figure 8. DREM path 3-2 regulated by FurB
Figure 9. Rv0081 activation/repression ratios for each DREM path during hypoxia
Figure 10. Macrophage and hypoxia data sets: comparison of time courses
Figure 11. DREM predictions for DosR and KstR as key regulators during the early stages of both hypoxia and macrophage models
Figure 12. Clustergram of the group of genes which has a different behavior in hypoxia and macrophage models
Figure 13. Genes involved in complex lipid biosynthesis and the upregulated DREM paths in the macrophage model that they are associated with
Figure 14. PhoP and WhiB3 time course expression and FFL

x
List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TB</td>
<td>Tuberculosis</td>
</tr>
<tr>
<td>MTB</td>
<td><em>Mycobacterium tuberculosis</em></td>
</tr>
<tr>
<td>TF</td>
<td>Transcription Factor</td>
</tr>
<tr>
<td>TRN</td>
<td>Transcription Regulatory Network</td>
</tr>
<tr>
<td>PAT</td>
<td>Polyaclytrehaloses</td>
</tr>
<tr>
<td>DAT</td>
<td>Diaclytrehaloses</td>
</tr>
<tr>
<td>PDIM</td>
<td>Phthiocerol Dimycocerosates</td>
</tr>
<tr>
<td>SL-1</td>
<td>Sulfolipid-1</td>
</tr>
<tr>
<td>TAG</td>
<td>Triacylglycerol</td>
</tr>
<tr>
<td>TMM</td>
<td>Trehalose Monomycolate</td>
</tr>
<tr>
<td>TDM</td>
<td>Trehalose Dimycolate</td>
</tr>
<tr>
<td>ChIP</td>
<td>Chromatin Immunoprecipitation Followed by Sequencing</td>
</tr>
<tr>
<td>iNOS</td>
<td>Nitric Oxide Synthase</td>
</tr>
<tr>
<td>ROI</td>
<td>Reactive Oxygen Intermediates</td>
</tr>
<tr>
<td>RNI</td>
<td>Reactive Nitrogen Intermediates</td>
</tr>
</tbody>
</table>
Chapter 1. Introduction

1.1. Mycobacterium tuberculosis (MTB) – the causative agent for tuberculosis (TB)

*Mycobacterium tuberculosis* (MTB) still remains a pathogen that causes millions of deaths each year [1-5]. The bacterium is primarily transmitted via the respiratory route. Infection occurs in the lungs but the organism can occupy any organ via hematogenous spread. Being exposed to MTB, 30% of individuals become infected, with about 40% of these individuals developing primary active tuberculosis (TB) and 60% developing latent infection. 2–23% of immunocompetent patients with latent TB will reactivate at a later date, while patients with HIV develop reactivation TB at a rate of around 5-10% per year [1]. Part of what makes MTB such a successful pathogen is its ability to survive within the host for months to decades in an asymptomatic latent state and to adapt to a wide range of conditions inside and outside the host.

1.2. Latency, dormancy and persistence

Three terms are commonly used when describing MTB and TB pathogenesis: latency, dormancy and persistence [5]. Latency was defined by Amberson as “the presence of any TB lesion which fails to produce symptoms of its presence” or can be also referred as the disease itself in its asymptomatic state [1, 4]. Without antibiotic treatment, chronic or latent infection is thought to be the typical outcome of TB infection. Latent TB can reactivate after years or even decades of subclinical persistence, leading to progressive disease and active transmission of the pathogen.

Dormancy is used to describe both TB as a disease as well as the metabolic state of the pathogen. TB lesions are described as active if the associated pathology is
progressing and dormant if it's healing. Active lesions are characterized by easily detectable populations of acid-fast, culturable MTB. The term dormancy is also often associated with an \textit{in vitro} model of MTB growth under hypoxic (oxygen-depleted) conditions developed first by Wayne [4]. This model is thought to approximate the state of MTB surviving in closed, necrotic lesions during clinical latency.

The term persistence literally means “continuing to survive steadfastly under the immune response of the organism and adverse conditions”. MTB, as a pathogen, can persist on several levels. On the cellular level, mycobacteria reside within macrophages, cells which eliminate pathogens and other foreign material from the body. At a more systematic level, MTB is able to avoid being eliminated from the human host despite the development of a vigorous immune response. Another important form of MTB persistence is the slow rate at which the bacterium is cleared by anti-TB drugs.

1.3. Modeling persistence in animals

The most commonly used model for studying TB is the mouse. Once delivered, MTB grows relatively unimpeded in the lungs (and spleen) of mice for the first 2 to 4 weeks after which the immune system controls the growth of the bacteria and their numbers reach a plateau. The immune response is not able to substantially reduce the bacterial numbers in the lung or spleen and as a result a persistent or chronic infection ensues and is maintained for many months. Despite relatively high levels of bacteria in the lungs, the mice do not show clinical signs of disease and can survive for more than a year postinfection. This is an attractive model as it represents
equilibrium between host and bacillus and the bacterial numbers are controlled by
the immune response, which holds for the human latent infection [3, 5, 22].

Another mouse latent TB model, so called “Cornell model”, involved treating MTB
infected mice with antimycobacterial drugs, which reduced the bacterial burden to
undetectable levels. This model is attractive because of that low bacterial burden.
But introduction of antibiotics to reduce the bacterial numbers does not mimic the
situation in natural human latent TB and may affect the development of a protective
immune response. However, some groups have used variations of this model to
study persistence and reactivation of mycobacteria in the host, as well as vaccination
of latently infected hosts [5].

Other animal models for the study of TB include guinea pigs, rabbits and non-human
primates. The rabbit model is the closest to imitating human TB in terms of tissue
pathology and disease progression. Still, rabbit as well as the other animal models
are not widely used compared to the mouse model [5, 22].

1.4. Macrophage infection

MTB is able to survive within the host and cause disease because of its remarkable
capacity to survive within the hostile environment of the macrophage [9, 10].

1.4.1. Formation of the granuloma

Once the infectious bacilli are inhaled as droplets from the atmosphere and reach the
lung, they are phagocytosed by alveolar macrophages and induce a localized
proinflammatory response that leads to the recruitment of mononuclear cells from
neighboring blood vessels. These cells form the granuloma, or, tubercle, which is the
signature structure of TB. The granuloma consists of infected macrophages surrounded by foamy macrophages and other mononuclear phagocytes. The periphery of the structure is mantled by lymphocytes associated with a fibrous cuff of collagen and other extracellular matrix components. This tissue response is typical for the 'containment' phase of the infection in which there are no overt signs of disease and the host does not transmit the infection to others. In the later stages, the granuloma develops a marked fibrous sheath. Containment usually fails when the immune status of the host changes, which is usually cause by old age, malnutrition or co-infection with HIV (any condition that impairs the function of CD4+ T cells). As a result, the granuloma caseates (decays into a structureless mass of cellular debris), ruptures and spills thousands of viable, infectious bacilli into the airways [9].

Both activated and non-activated macrophages coexist within granulomas, with the activated macrophages processing and representing mycobacterial antigens to the surrounding lymphocytes. As for MTB itself, the precise status of the bacteria within granulomas is not clear. They may exist as actively dividing bacilli or in a “dormant” state [10]. And it’s possible that both actively dividing and dormant bacilli can occur within the same infected individual depending on the stage of the disease [14, 15].

1.4.2. Mycobacterial prevention of phagosome-lysosome fusion

MTB is able to survive within the phagosome by arresting its normal maturation. Instead to pH 5 the phagosomes containing the bacteria acidify only to pH 6.4 limiting the fusion with the pre-formed lysosomes [9, 10]. The modulation of the phagosome seems to be mediated by both cell-wall lipids and other bacterial effectors [9]. MTB has a thick cell wall consisting of unique lipid and glycolipids
moieties. The glycolipids of the bacteria can interfere with the phagosome fusion through blocking a normal host trafficking event that is regulated by phosphatidylinositol 3-phosphate (PIP3). PI3P is a host membrane component that is essential for phago-lysosome biosynthesis [26]. PIP3 is thought to present a docking site for several proteins involved in the maturation of the phagosomes into lysosomes [10]. MTB interferes with this trafficking event by preventing PIP3 accumulation on phagosomal membranes [27]. There are two strategies used for the process. First, MTB interferes with the activity of the PI3 kinase hVP34, preventing the generation of PIP3 on the phagosomal membrane and blocking phagosome-lysosome fusion. This inhibitory activity was proposed to be mediated by the mycobacterial cell-wall component lipoarabinomannan (LAM) since the addition of purified LAM was sufficient to prevent lysosomal delivery [28]. It remains unclear how LAM-mediated blockage is regulated. A second strategy relies on the prevention of PIP3 accumulation on phagosomes by the activity of a phosphatase, termed SapM [29]. SapM is presumably released within the host cell cytosol upon infection, where it may hydrolyze PIP3 on phagosomal membranes [30]. Apart from lipid phosphatases, MTB also possesses protein phosphatases (PtpA and B [31]) that interfere with host trafficking processes.

One of MTB serine/threonine kinases, PknG, is essential for the prevention of phagosome-lysosome fusion, as the overexpression of a mutant form of PknG results in lysosomal delivery of the bacteria [32]. As several factors that are involved in the regulation of intracellular transport reactions require phosphorylation, it is likely that PknG acts by phosphorylating a host molecule, thus preventing the activity of this host factor in carrying out phagosome-lysosome fusion.
Apart from producing virulence factors such as SapM and PknG, MTB has evolved a mechanism to interfere with host cell signaling. A protein, referred to as coronin 1 (or TACO) is recruited to phagosomes containing live bacilli, but rapidly released from phagosomes containing killed mycobacteria, which suggests that coronin 1 is an important host factor that specifically prevents the lysosomal delivery and death of the bacilli inside macrophages [33, 34]. The protein prevents phagosome-lysosome fusion by regulating calcium-dependent signaling processes. When macrophages are infected with MTB, a sustained calcium flux is generated that is dependent of the presence of coronin 1. The coronin 1-dependent cytosolic influx of calcium activates the calcium-dependent phosphatase calcineurin. The activation of calcineurin is required for blocking phagosome-lysosome fusion [35]. The activated protein links up to several downstream signaling cascades, ranging from the transfer of transcriptional factors to the nucleus to dephosphorylation of proteins involved in endocytosis [36, 37]. The mechanism of coronin 1 recruitment to the mycobacterial phagosome and the precise activity of calcineurin required to prevent lysosomal delivery of mycobacteria are unknown.

Mycobacterial glycolipids also play an important role in the subversion of macrophage activation. The abundant cell-wall component LAM as well as its glycosylated forms (manLAM) can modulate signaling pathways that induce macrophage activation, including interferon-γ-mediated gene expression, TLR activation, and phagosome-lysosome fusion [38]. LAM can also block the activation of an important kinase (MAPK) that is downstream of various stimuli that cause macrophage activation [39].
In addition, modulation of bacterial metabolic pathways can lead to a long-term survival within macrophages. When bacteria enter a nutrient-poor environment similar to the one that can occur within macrophage phagosomes, they become restricted to fatty acids as their carbon source. An activation of the glyoxylate shunt pathway follows, which facilitates carbon retention through de novo synthesis of carbohydrate [40]. The first step in this pathway is catalyzed by isocitrate lyase (icl1) which is an essential enzyme for the establishment of a persistent infection of MTB.

Another cell envelope component that contributes to acid resistance and virulence of MTB is the outer membrane protein OmpATb [41, 42]. An OmpATb mutant showed delayed growth at pH 5.5 in vitro, and reduced growth in macrophages and mice. The mechanism by which OmpATb confers acid resistance is unknown. In lipid bilayers this protein exhibits pore-forming activity, which is pH-sensitive as the channel has a propensity to close at low pH [42].

Two other genes that may have a role in acid resistance are Rv3671c and Rv2136c [41]. Their transposon mutants were impaired in their ability to maintain a neutral intrabacterial pH during acid challenge and inside activated mouse macrophages [43]. Rv3671c encodes a membrane-associated serine protease and might protect MTB against acid by modifying the bacterial cell envelope, regulating protein or lipid quality control and/or serving in signaling pathways that help the bacteria to resist extracellular stress. Rv2136c is a putative homologue of Escherichia coli’s UppP, an undecaprenol pyrophosphate phosphatase involved in peptidoglycan biosynthesis [41]. The exact mechanisms by which these two genes protect against acid stress and support virulence is still unknown.
Mg$^{2+}$ transport is another potential stress-resistance factor as an $MTB\Delta mgtC$ mutant was attenuated for growth in macrophages and mice. MgtC is a putative Mg$^{2+}$ transporter and Mg$^{2+}$ might be required at low pH for the maintenance of cell wall integrity, as a cofactor for enzymes that become important during acid stress or for the function of Mg$^{2+}$-dependent ATPase involved in extruding cytosolic protons [44].

1.4.3. Systems for regulation of MTB metabolic and growth state in the granuloma

MTB possesses at least three systems that regulate its metabolic and growth state in latent phase [16]. First, a well-characterized bacterial regulon that is controlled by DosR-DosS – a two-component system in mycobacteria – is induced by several stimuli, thought to play role in the latent TB, including local hypoxia [17, 18], nitric oxide [19] and carbon monoxide [20, 21]. This ‘dormancy’ regulon controls the expression of genes that allow the bacteria to use alternative energy sources, in particular lipids, and genes encoding factors that are selectively recognized by T cells from humans with latent TB. Second, MTB has a system of five proteins that resemble the well-characterized Micrococcus luteus resuscitation-promoting factor (Rfp), which is a secreted protein that has the ability to ‘resuscitate’ bacteria from a nutrient-starved dormant state [22]. Deletion of one or more of MTB Rfp genes generates bacteria that have an impaired recovery from dormancy, indicating that these genes may participate in the progression from latency to reactivation [23, 24]. Finally, MTB encodes 88 toxin-antitoxin gene pairs, the expression balance of which regulates multiple phenomena, including whether the bacteria replicate or remain static [25].
1.4.4. MTB defense mechanisms against oxidative and nitrosative stress

Activated macrophages express two enzymes, phagocyte oxidase (NOX2) and inducible nitric oxide synthase (iNOS) which generate reactive oxygen intermediates (ROI) and reactive nitrogen intermediates (RNI) respectively [41]. Upon phagocytosis, the preformed NOX2 subunits assemble into an enzymatically active enzyme complex that transfers electrons across the membrane from cytosolic NADPH to molecular oxygen. As a result, superoxide anions are produced. They dismutate into hydrogen peroxide (H$_2$O$_2$) and generate toxic hydroxyl radicals [45]. iNOS is induced by IFNγ and produces nitrite and nitrate via nitric oxide. Under the acidic conditions of the phagosomes of IFNγ-activated macrophages, nitrite forms nitrous acid, which dismutates to nitric oxide (NO) and another toxic radical, nitrogen dioxide [46]. NO can synergize with superoxide (produced by macrophages or generated as a byproduct of respiratory metabolism by the pathogen) to form poisonous peroxynitrite [47]. These ROI and RNI can be bactericidal as they react with a wide variety of molecules, including nucleic acids, proteins, lipids and carbohydrates. To counteract such stress, MTB has a variety of molecules to either detoxify ROI and RNI before they can harm the bacteria or to repair damage which they caused.

One such resistance mechanism is mediated by the katG product catalase-peroxidase, which decomposes H$_2$O$_2$ into water and oxygen. MTB lacking katG exhibited no catalase activity and was hypersusceptible to H$_2$O$_2$ in culture [48]. MTB contains two genes encoding superoxide dismutases – enzymes catalyzing the conversion of superoxide anions to hydrogen peroxide – sodA and sodC. SodC is a
Cu,Zn superoxide dismutase localized to the mycobacterial cell envelope [49]. Lack of sodC increases susceptibility of MTB to superoxide, to the combination of superoxide and nitric oxide and to killing by IFNγ-activated murine macrophages [50]. In contrast, resting wild type macrophages did not kill the sodC mutant which suggests that Cu,Zn superoxide dismutase contributes to MTB’s resistance against oxidative burst products generated by activated macrophages. Its role during the long-term chronic infections has not been determined. SodA, which uses iron instead of Cu,Zn, may compensate for SodC to protect against the respiratory burst in resting macrophages and during mouse infections. SodA appears to be essential for in vitro growth [51]. MTB mutants with reduced sodA expression displayed increased susceptibility to H2O2 and were significantly attenuated in mice [52].

5'-adenosinephosphosulfate reductase CysH also has antioxidant activity as its mutant displayed increased susceptibility to H2O2 and peroxynitrite [53]. Decreased levels of downstream metabolic products including mycothiol could be the cause for this.

To detoxify both RNI and ROI, MTB expresses an NADH-dependent peroxydase and peroxynitrite reductase that consists of four proteins: alkylhydroperoxide reductase subunit C (AhpC), a thioredoxin-related oxidoreductase (AhpD), dihydrolipoamide acyltransferase (DlaT) and lipoamide dehydrogenase (Lpd) [54, 55].

A novel MTB gene, noxR1, has been shown to confer resistance to the toxic effects of RNI, although the precise mechanism is not known. The noxR3 gene also protected bacteria from ROI and RNI [56].
The sulfur of cysteine and methionine residues in proteins is highly susceptible to oxidation and exposure of methionine to ROI and RNI leads to the generation of methionine sulfoxide, which can interfere with protein function. Methionine sulfoxide can be converted back to methionine by methionine sulfoxide reductase (Msr). Msr might also serve as a sink for oxidants before they damage cellular macromolecules. MTB expresses two Msr enzymes. When lacking both msrA and msrB, the bacteria was more readily killed by acidified nitrite, but the mutant was not hypersusceptible to $\text{H}_2\text{O}_2$ [41].

It has been also shown that mutations in the excision repair gene uvrB result in severe sensitivity to RNI [41].

1.4.5. Difference between murine and human macrophages

There is a controversy about the differences between rodent and human macrophages [56-58]. Some researchers have even suggested that human macrophages are “fundamentally” different from rodent macrophages [57]. The central problem of the controversy is the use of arginine for NO production. As mentioned above, in macrophages NO is produced by iNOS, an oxidoreductase that catalyzes the conversion of arginine and oxygen to NO and citrulline. The NOS2 gene (encoding iNOS) is present in both rodents and humans. In mouse models of intracellular infection NO production by iNOS has been established as essential for effective immunity to MTB and several other bacterial species [58]. However, the role of NO in human macrophages is unexpectedly controversial. While in murine macrophages it has been shown that a significant amount of NO is produced, in human monocyte-macrophage cell lines no NO production has been identified.
Moreover, in direct opposition of the *in vitro* data, human tissues, such as TB biopsy samples, express abundant iNOS. What can be the reason for this discrepancy? Experiments in human macrophages to measure iNOS and NO are performed almost universally in cells derived from donor monocytes and are “differentiated” *in vitro* in media such as RPMI. There is no guarantee that such a system is permissive to iNOS expression. The simplest hypothesis is that the culture conditions and/or the cell type are lacking in one or more factors that allow iNOS expression in human macrophages *in vitro* [58].

A similar scenario exists for the arginases. There are two arginase enzymes: Arg1 and the mitochondrial Arg2. In humans Arg1 appears to be expressed only in neutrophils, and mouse Arg1 is expressed in macrophages and neutrophils. Mouse Arg 2 is also expressed in macrophages and neutrophils [58].

1.5. Modeling persistence *in vitro*

The stationary phase culture is the simplest model of MTB persistence. The kinetics of replication of the bacteria in the lungs of mice are similar to the organism’s replication in culture where an initial period of exponential growth is followed by an extended period in which the number of viable bacilli remains stable. As a result, the stationary phase culture is a simple, inexpensive, and easily manipulated system to analyze the long-term survival of non-replicating cells of MTB. Important observations about MTB drug susceptibility and response were made using this type of culture [5].

A well-known *in vitro* model of dormancy was developed by Wayne [6, 59]. The system relies on a self-generated O₂ gradient. In the model a sealed, standing
culture is allowed to incubate over a period of days while the bacteria deplete the available oxygen. The model demonstrated that MTB was capable to survive in anaerobic conditions (supposedly the conditions that the bacteria face in lesions) in vitro and the bacteria that survived anaerobiosis underwent stage-specific cell cycle arrest represented by two non-replicating persistence states. The dormant state induced by oxygen deprivation (hypoxia) was characterized by almost total shutdown of DNA, RNA and protein synthesis and relative resistance to antimycobacterial agents such as isoniazid. Thus, adaptations to hypoxia are thought to play a prominent role in MTB pathogenesis and latency.

Despite improvements, the Wayne model still has the issue of reproducibility. A defined hypoxia model is an alternative in vitro model where the bacilli are exposed to defined hypoxic atmospheres that remain constant throughout the experiment. In this approach, oxygen tensions of 1% or less halt replication but bacteria remain viable [60]. A defined hypoxic model has been used to obtain the transcriptomics data which will be introduced in chapter 2.

1.6. Signature genes involved in hypoxic response

Wayne’s in vitro model of MTB latency has been used to demonstrate the upregulation of the glyoxylate cycle, a metabolic pathway that allows the bacteria to use acetate or fatty acids as the sole carbon source. The activity of a key enzyme of the cycle, isocitrate lyase (icl) has been shown to increase in long-term cultures of MTB [7], and the disruption of icl attenuated the ability of MTB to persist in a murine model of infection [8]. Activation of infected macrophages increased expression of
*icl*, and the *icl* mutant was attenuated for survival in activated but not resting macrophages.

Nitrate reductase activity is greatly accelerated during the first nonreplicating state in Wayne’s model [4]. Although MTB contains genes for nitrite reductase, reduction of nitrite has not been demonstrated and ammonium ion does not inhibit the reduction of nitrate. Thus, nitrate reductase probably serves a mainly respiratory, rather than assimilatory function, providing maintenance energy during non-replicating persistence. Two loci represent the nitrate reductase activity in the bacteria. The first locus is *narGHJI* which was shown to have reductase activity [12]. The second locus is called *narX* which was shown to be upregulated during anaerobic conditions in BCG [13].

Another gene, alpha crystallin homologue - *hspX*, a chaperone protein, has been reported to be produced by *M. tuberculosis* only when the bacilli encounter hypoxic conditions or are exposed to nitric oxide but not to a variety of other stresses, including heat shock [11].

1.7. Transcriptional regulation in latent phase

1.7.1. DosR

DosR (Rv3133c) is a transcriptional regulator involved in the initial response to hypoxia [61, 64]. Phosphorylation of DosR by either of two sensor histidine kinases, DosS or DosT [62], leads to the induction of a set of ~50 genes, many of them with unknown function. Among the genes with known or predicted function there are several that may play a role in adaptation to hypoxic stress: *acr* (stabilizing partially denatured proteins); *narX*, *narK2* and *fdxA* (nitrate accumulation and alternative
electron transport); *nrdZ* (dNTP synthesis under microaerophilic conditions); *tgs1* (triglyceride synthase) and six MTB orthologues of the universal stress protein family (resistance to DNA damage) [59]. It has been shown that the DosR-dependent regulon is also induced in response to nitric oxide [65], in macrophages [66], in both early [67] and late mouse infections [68], and in response to carbon monoxide [69], SDS [70] and low pH [71]. Non-replicating persistence *in vitro* and chronic infection *in vivo* occur despite disruption of the DosR regulon suggesting that rather than in dormancy response, the DosR regulon may play a much more important role in survival of extended respiratory, nitrosative or redox stress [59]. Transcriptional analysis under hypoxic conditions *in vitro* of an extended time course showed the existence of an additional so-called Enduring Hypoxic Response (EHR) comprising a set of 230 genes [63]. The number of transcription factors (TFs) in EHR regulon is higher than expected suggesting a complex regulatory circuit. These TFs include the iron uptake regulators FurA and FurB, three members of the WhiB family, the two-component response sensor PhoP. Alternative sigma factors SigE and SigH are also induced during EHR [63].

1.7.2. KstR

KstR (Rv3574) is a member of the TetR family of transcriptional regulators that controls the expression of a large regulon involved in lipid metabolism in mycobacteria [72]. There are 10 *fad* genes (*fadA5, fadE26, fadE27, fadE28, fadE34, fadD8, fadD12, fadD17 and fadD19*), one *ech* gene (*echA19*), one *lip* gene (*lipO*), and at least three ketoacyl-CoA thiolases (*ltp2, ltp3* and *ltp4*). In addition, the *mce4* operon is part of this regulon, and it has been suggested that the *mce* operons are involved in the import of lipids or are lipid-associated [73]. Other genes in the KstR
regulon, \textit{hsaC} and \textit{hsaD} (formerly \textit{bphC} and \textit{bphD} respectively), have been implicated in both cell wall synthesis [74] and cholesterol degradation [73]. Later studies for these two genes and some others have confirmed a role in cholesterol catabolism [73]. A second TetR-type repressor, KstR2, controls the expression of a small regulon (15 genes) which may also play a role in cholesterol utilization in mycobacteria. KstR and KstR2 coordinately control more than 70 genes that are all de-repressed by growth on cholesterol [74].

1.7.3. PhoP

PhoP (Rv0757) is a transcriptional regulator of the two-component system PhoPR [76-78]. Inactivation of \textit{phoP} leads to high attenuation of MTB. However, it is not completely eliminated and persists \textit{in vitro} cultured-macrophages and in mouse organs [79]. Further supporting the role of PhoP in virulence regulation, it has been shown that a point mutation in PhoP contributes to avirulence of the H$_{37}$Ra strain, since this mutation disrupts secretion of the ESAT-6 antigen and the synthesis of acyltrehalose-based lipids in this strain [80-82]. Comparing both the transcriptome and the proteome of MTB wild type with a \textit{phoP} mutant, another group has extended the regulatory network of PhoP and assigned it to several functional groups [77]. PhoP controls some of the genes of the DosR regulon including \textit{dosRS} genes themselves. Alpha crystallin, a latency antigen, also appeared downregulated in the \textit{phoP} mutant. Altogether, these observations indicate that PhoP might regulate partially the dormancy regulon through cross-talking to DosR. PhoP also regulates a subset of genes from EHR.
A second group of genes regulated by PhoP have respiratory functions. PhoP positively regulates \textit{nuo} genes from the NADH dehydrogenase operon. This enzymatic complex functions as a primary electron acceptor via oxidation of NADH to NAD$^+$. NuoG inhibits apoptosis in macrophages and increases virulence in immuno-compromised mice [83]. PhoP also regulates the expression of the enzyme alanine dehydrogenase (\textit{ald}). This enzyme participates in maintaining the NADH pool by recycling NAD$^+$ through the conversion of pyruvate to alanine when oxygen, as a terminal electron receptor, becomes limited [84]. Additionally, PhoP controls the genes involved in utilization of nitrogen and sulfur sources in oxygen limiting conditions such as the nitrite transporter \textit{narK1} and the sulfur reduction operon \textit{nirA-cysH} [85].

PhoP has been shown to positively regulate genes within the RD1 genomic region. The RD1 region is essential for MTB virulence and encodes the secretion system ESX-1 which assures export of the major T-cell antigen complex ESAT-6/CFP10 [86-88]. Potential transcriptional mechanism for control of ESAT-6 secretion involves \textit{espB} gene and the ESX-1 secreted protein regulator \textit{espR} both of which are regulated by PhoP [77].

PhoP positively regulates genes implicated in MTB lipid metabolism. These genes include \textit{pks3}, Rv1184c, \textit{fadD21} and \textit{pks2}, involved in the synthesis of acyltrehalose-based lipids [89], \textit{lipF} coding for a lipid esterase required for virulence in mice [90], and \textit{fadD9} which encodes a hypothetical fatty acid-CoA ligase. Additionally, PhoP controls the fatty acids gene encoding a fatty acid synthase which together with the FAS II system generates precursors for the synthesis of mycolic acids.
All the functional groups regulated by PhoP are also differentially expressed in response to macrophage infection indicating that PhoP might control key functions for intracellular survival [77].

PhoP has been shown to positively regulate the aprABC locus whose gene expression is induced during growth in acidic environments in vitro and in macrophages [78]. Deletion of this locus causes defects in gene expression that affect aggregation, intracellular growth, and the relative levels of storage and cell wall lipids. A model has been proposed where phoPR senses the acidic pH of the phagosome and induces aprABC expression to change the expression of the set of genes this locus regulates. Additionally, there is a positive feedback loop of aprABC on phoP which should enable rapid adaptation for the host environment [78].

1.7.4. WhiB3

WhiB3 (Rv3614) is a high-level redox regulator in MTB [91-93]. WhiB3 senses fluctuations in the intracellular environment associated with O₂ depletion, including the metabolic switch to the favored in vivo alternative carbon source, fatty acids [92]. WhiB3 has been implicated in several metabolic functions. First, this TF is essential for maintaining bacterial shape and size [91, 92]. Second, WhiB3 regulates the anabolism of MTB complex virulence lipids such as polyacyltrehaloses (PAT), diacyltrehaloses (DAT), sulfolipids (SL-1), and phthiocerol dimycocerosates (PDIM), which involves the use of large quantities of reducing equivalents (NAD[P]H). Third, WhiB3 was shown to maintain redox homeostasis during infection of macrophages [92]. Fourth, the MTB ΔwhiB3 showed decreased production of PAT, DAT and SL-1, but increased production of PDIM, and to a lesser extent triacylglycerol (TAG) [92].
MTB ΔwhiB3 accumulates PDIM and TAG during intramacrophage growth, suggesting that this intermediate is channeled into PDIM via the methyl-malonyl CoA (MMCoA) pathway and into TAG as a result of the increased resistance to propionate toxicity [93]. As mentioned above, PhoP also regulates positively the production of PAT, DAT and SL-1 but the expression of phoP is not altered in MTB ΔwhiB3 [92] and MTB ΔphoP did not accumulate PDIM or TAG. Thus, the WhiB3-dependent control of cellular redox homeostasis may be an additional factor that is required to regulate the flux of propionyl CoA to methyl-branched polyketides and TAG synthesis.

WhiB3 was shown to bind to the upstream sequence of lipid biosynthesis genes pks2 and pks3 [92, 93].

1.7.5. Sigma factors

MTB encodes a repertoire of 13 σ factors [94]. SigA is indispensable for growth in MTB. The sigA transcript is maintained at a constant level under various stress conditions. For that reason sigA continues to be widely used as an internal standard for normalization in quantitative RT-PCR experiments. A point mutation in one of the domains of SigA resulted in attenuation of virulence of M. bovis strain ATCC 35721 in a guinea pig model of infection. Overexpression of sigA in MTB H₃₇Rv enhanced its growth in human macrophages and in the lungs of mice after aerosol infection, which further suggests its role in virulence [94].

SigB has been found to be dispensable for growth in M. smegmatis and MTB. The expression of sigB increases upon exposure to different environmental stresses such as low aeration, treatment with hydrogen peroxide and heat shock. The inactivation
of the *sigB* gene does not affect the survival of MTB during infection in human macrophages or in mouse and guinea pig models. However, deletion of *sigB* in MTB results in its higher sensitivity to SDS-induced surface stress, heat shock, oxidative stress, exposure to vancomycin and hypoxic conditions. Overexpression of *sigB* resulted in upregulation of genes encoding proteins involved in cell wall-related processes, ESAT-6-like proteins, 50S ribosomal proteins, PE-PGRS proteins, keto-acyl synthase, KasA and the TFs WhiB2, IdeR and SigB itself. Expression profiling of MTB *sigB* mutant showed regulation of *ideR*, *furA*, *katG*, *ppe19* and *hsp20* by SigB [94].

SigC is downregulated during stationary phase and in response to heat shock and SDS-induced surface stress [94]. SigC is not required for the survival of MTB in murine bone marrow-derived macrophages. In a mouse model of infection, *sigC* mutant in both CDC1551 and RV37Rv backgrounds grows and persists in lungs but shows attenuated disease progression. The attenuation of *sigC* mutant may result from dysregulation of expression of several key virulence-associated genes, such as *hspX*, *senX3*, *mtrA*, polyketide synthases and *fbpC* (antigen 85C).

SigD is expressed at a moderately high and constitutive level during exponential and stationary growth phases and declines significantly, following hypoxia, a pattern very similar to that of SigA in an *in vitro* culture. This response is modulated by the *relA* gene product via the synthesis of a key signaling molecule, guanidine tetraphosphate (ppGpp). *RelA* deletion in MTB results in a loss of virulence accompanied with a significant decrease in *sigD* expression during logarithmic phase, *sigD* is upregulated during nutrient starvation, further suggesting the role of SigD in physiological
adaptations and starvation. Some of the important genes regulated by SigD include those encoding proteins involved in lipid metabolism, cell wall-related processes, stress response and DNA binding and repair. Several genes, such as *rpfC* (implicated in the revival of dormant bacteria), *pk*10 (polyketide-like chalcone synthase), *recR* and those encoding several chaperones, ribosomal proteins, elongation factors and ATP synthase subunits were also reported to be downregulated in the MTB *sigD* mutant [94].

SigE is upregulated in mycobacteria grown in human macrophages compared with those grown in an *in vitro* culture. Its expression also increases under exposure to heat shock, SDS-mediated surface stress, isoniazid and vancomycin [94]. The MTB *sigE* mutant showed decreased viability in macrophages. SigE-dependent genes encode proteins that belong to different classes such as TFs (including SigB, Rv3050c, MprA and MprB), enzymes involved in fatty acid metabolism (most importantly icl) and the classical heat shock proteins [94].

MTB H₃₇Rv *sigF* was found to be upregulated in a nutrient starvation model of MTB and during infection of cultured human macrophages. However, it was shown that SigF is not required for MTB survival under nutrient starvation conditions and within activated murine macrophages or for extracellular persistence in an in vivo granuloma model of latent TB infection [94]. SigF regulates the expression of genes involved in the biosynthesis and structure of the mycobacterial cell envelope, including complex polysaccharides and lipids, particularly virulence-related sulfolipids. SigF also regulates a number of TFs including PhoY1 and Rv2884 as well as SigC. Conditional overexpression of *sigF* during the early exponential phase
resulted in upregulation of several genes encoding PE and PPE family proteins and mmpL family transporters known to be involved in virulence.

*sigG* is one of the most highly induced genes in MTB during macrophage infection, and has been shown to be required for the survival of the bacteria. However, its expression is downregulated upon exposure to various stress conditions, such as mild shock, heat shock, low aeration and SDS-mediated surface stress. The *sigG* mutant strain showed downregulation of several genes encoding proteins involved in fatty acid metabolism such as *AceA* (isocitrate lyase), *FadE5* (acyl-coenzyme A dehydrogenase) and *ScoA* (succinyl-coenzyme A) [94].

SigH is a central regulator of oxidative and heat stress responses in MTB. Although the expression of *sigH* was found to be induced during macrophage infection, the *sigH* mutant was not attenuated for growth in human macrophages. The SigH regulon includes its own structural gene and genes encoding SigB, SigE, Rv0142 (putative transcriptional regulator), DnaK, ClpB (heat shock proteins), TrxB and TrxC (thioredoxin reductase/thioredoxin). SigH also induces enzymes involved in cysteine biosynthesis and in the metabolism of ribose and glucose [94].

Unlike *sigE* and *sigH*, *sigL* does not play a role in oxidative and nitrosative stress responses. Its regulon includes genes involved in lipid biosynthesis (*pks7, pks10, ppsA*), fatty acid transport (*mmpL13a, mmpL13b, sigB* [94].

Overexpression of SigM revealed upregulation of four putative esat-6 homologs (*esxT, esxU, esxE* and *esxF*). Positive regulation of Esx family secretory proteins by SigM suggests its role in long-term adaptation to specific host environments. Other
genes identified to be regulated by SigM include PPE family members \((ppe1, ppe19, ppe60)\), fatty acids, \(pks2\) and \(pks3\), genes from the first operon of PDIM biosynthetic and transport locus, \(mas\) gene, \(kasA-kasB\) operon, required for mycolic acid synthesis \[94\].

1.8. Lipid metabolism

The metabolism of host lipids and the production of cell wall lipids are central for the survival of intracellular MTB. There is growing body of evidence supporting the fact that the bacterium is able to utilize host lipids, such as fatty acids and cholesterol, to persist within the host cell. Host lipids appear to be the primary carbon source for MTB in vivo or in infected macrophages. MTB was shown to have the incredible ability to simultaneously co-catabolize multiple carbon substrates \[99\]. Moreover, during co-catabolism, the bacterium catabolizes each carbon source differentially through the glycolitic, pentose phosphate and/or tricarboxylic acid pathways to distinct metabolic fates. In total, the lipid metabolic machinery of MTB consists of about 250 distinct enzymes compared to only 50 in \(E. coli\) \[95\].

1.8.1. Cholesterol metabolism

Cholesterol is known to be one of the main host lipids that are required for the growth and persistence of MTB in mice \[104\], and mutants defective in either uptake or processing of cholesterol exhibit severe defects in intracellular growth and survival \[105-107, 110-111\]. Studies have shown that cholesterol is essential for the uptake of mycobacteria by macrophages, and it has been found to accumulate at the site TB entry and mediate the phagosomal association of TACO, a coat protein that prevents degradation of mycobacteria in lysosomes \[106\]. Bioinformatic analysis in
Rhodococcus sp. [108] and subsequent more recent mutagenesis studies in MTB [111] identified 83 genes that are essential for MTB growth *in vitro* and cholesterol catabolism. These genes are categorized in three functional groups associated with cholesterol utilization: (1) steroid ring degradation, (2) side-chain degradation, and (3) intermediary metabolism [112]. Cholesterol has been shown to be utilized via androstenedione/androstadienedione pathway and this process requires production of an intact KstD enzyme [110]. The growth of MTB on cholesterol was significantly affected by a knockout of the *mce4* gene, which encodes an ABC transporter responsible for cholesterol uptake [104]. Another locus, *igr*, which was previously found to be essential for intracellular growth and virulence of MTB, is required for cholesterol metabolism: *igr*-deficient bacteria cannot grow using cholesterol as a primary carbon source [105]. Recently, it has been demonstrated that growth on cholesterol requires the transcriptional induction of the propionyl-CoA-assimilating methylcitrate cycle enzymes via the Rv1129c regulatory protein [109]. Additionally, FadD3, regulated by Kstr2, was shown to be an acyl-CoA synthetase that initiates catabolism of cholesterol rings C and D [113], and FadA5 was found to be a thiolase required for virulence and production of androstenedione and androstadienedione [113].

1.8.2. Fatty acid metabolism

Fatty acids are the other carbon source utilized by MTB during infection. Evidence dates back to Bloch and Segal [100], who reported the preferential utilization of fatty acids as carbon source by MTB in the lungs of infected mice. When the MTB genome was sequenced, another group [95] showed the relative “overrepresentation” of genes predicted to be involved in the processing and
degradation of fatty acids. Additionally, genes involved in routing the products of fatty acid degradation through pathways such as the glyoxylate shunt and gluconeogenesis significantly influence the bacterium’s phenotype in a range of different infection models [8, 100-103]. The genes encoding fatty acid β-oxidation enzymes together with those of the glyoxylate cycle are shown to be up-regulated during infection in macrophages [66] and mice [96-97]. The source and types of fatty acids that might be utilized by MTB during infection are still unknown. One possible source is the complex long-chain fatty acids such as dipalmitoyl phosphatidylcholine, present in the lung surfactant, which the bacteria can metabolize in a glyoxalate cycle-dependent manner [98]. Fatty acids can be also derived by hydrolysis of lipids in the phagosomal membrane.

1.8.3. Mycolic acid metabolism

Mycolic acids are very long-chain fatty acids representing essential components of the mycobacterial cell wall [114]. They are found, primarily, as building blocks of arabinogalactan but also present as extractable “free” lipids within the cell wall, mainly associated to trehalose to form trehalose dimycolates (TDM). TDMs are the most abundant and significant toxic lipid extractable from the cell surface of virulent MTB. Mycolic acids provide MTB with resistance to chemical injury, resistance to dehydratation, low permeability to hydrophobic antibiotics, virulence and the ability to persist within the host. In addition, the enzymes involved in mycolate biosynthesis are essential for survival of the bacterium within the host. Free mycolic acids were observed to increase during hypoxia and be reversed in re-aeration, while opposite effects were found for trehalose monomycolates (TMMs) and TDMs [159].
The biosynthetic pathway of mycolates involves two types of fatty acid-synthesizing systems, the type I and type II fatty acid synthases, FAS-I and FAS-II, respectively. FAS-I catalyses the de novo synthesis of fatty acids from acetyl-CoA. FAS-II is composed of four dissociable enzymes that act subsequently and repeatedly to elongate the growing acyl chain. The biosynthetic process involves five stages: (1) the synthesis of the C\textsubscript{26} saturated straight chain fatty acids by FAS-I to provide the α-alkyl branch of the mycolic acids, (2) the synthesis of the C56 fatty acids by FAS-II to provide the meromycolate backbone, (3) the introduction of functional groups to the meromycolate chain by numerous cyclopropane synthases, (4) the condensation reaction catalyzed by the polyketide synthase Pks13 between the α-branch and the meromycolate chain before a final reduction to generate the mycolic acid and (5) the transfer of mycolic acids to arabinogalactan and other acceptors such as trehalose via the Antigen 85 complex.

The MTB fas gene (Rv2524c) encodes the multifunctional FAS-I polypeptide that contains all the functional domains required for de novo fatty acid synthesis. All intermediates generated remain enzyme-bound during the process of elongation and undergo transacylation to other catalytic sites within the enzyme. FAS-I generates short-chain fatty acyl-CoA primers that are further elongated by FAS-II.

Several enzymes are involved in the elongation process of FAS-II. MtFabH is the initiator of mycolic acid elongation and it catalyzes the condensation reaction of Acyl-CoA primers with malonyl-AcpM. MabA, FabG1 and InhA catalyze the next step of keto-reduction, dehydration and enoyl-reduction of the acyl-CoA AcpM precursor. KasA and KasB are two β-ketoacyl-AcpM synthases that catalyze the initiation of
subsequent rounds of acyl extension by FAS-II. KasA is suggested to catalyze the initial elongation reactions while KasB extends the elongation to full-length mycolates. KasB was found to be essential for MTB virulence. The final step in the synthesis of mycolic acids in MTB is the condensation of C26-S-CoA and meromycolyl-AMP. Pks13 (Rv3800), a member of the type I polyketide synthase family, has been shown to participate in this final step in mycolate assembly has been demonstrated. A specific fatty acyl-AMP ligase (FadD32) converts each meromycolyl-S-AcpM derived from the FAS-II system to meromycolyl-AMP. Following release from FAS-I, the C26-S-CoA is carboxylated by the acyl-CoA carboxylases AccD4 and AccD5 to yield 2-carboxyl-C26-CoA.

1.8.4. TAG accumulation and catabolism

Fatty acids are normally stored as TAG in the adipose tissues of mammals, seed oils of plants and as lipid inclusion bodies in prokaryotes for use as energy source during and after dormancy/hibernation. Based on that observation, TAG was postulated to be the storage form of energy for latent MTB [115]. Intracellular lipid inclusion bodies were initially observed in MTB more than fifty years ago and were more recently detected in mycobacteria isolated from the sputum of TB patients [116]. TAG accumulation was shown to be a critical event of MTB dormancy [117]. Triacylglycerol synthase 1 (Tgs1) was discovered to be the primary contributor to TAG synthesis within MTB and the deletion of tgs1 led to a nearly complete loss in TAG accumulation by MTB under in vitro dormancy-inducing conditions [118]. The source of fatty acids for synthesis of the TAG that accumulates as lipid droplets in the pathogen remains unknown. The lipid-loaded foamy macrophages which are found inside the hypoxic environment of the TB granuloma contain abundant stores
of TAG and are thought to provide a lipid-rich microenvironment for MTB. Human macrophages cultured under hypoxia (1% O2) accumulate TAG in lipid droplets [119].

Recently a group has demonstrated that MTB imports fatty acids released from host TAG to accumulate TAG within the bacterial cell [115]. Host fatty acids were incorporated intact into MTB TAG. They also show that host TAG is imported by MTB and accumulate as lipid droplets within the bacterial cell. Deletion of tgs1 resulted in a drastic decrease of TAG accumulation within MTB inside macrophages suggesting that synthesis of TAG within the pathogen from fatty acids released from host TAG constitutes the major pathway of TAG accumulation by the pathogen inside the host [115]. The accumulation of TAGs as well as high expression of tgs1 was also observed during hypoxia [159].

1.8.5. Complex lipid metabolism (SL-1, PAT/DAT, PDIM)
MTB possesses a complex cell wall that contributes to its pathogenesis and resistance to therapeutics. The cell wall is represented by multiple layers exterior to the cytoplasmic membrane and consists not only of peptidoglycan but also arabinogalactan polysaccharide layers and an extremely hydrophobic bilayers known as the mycobacterial outer membrane [120, 122] (Fig. 1). The base of the mycobacterial cell wall is the plasma membrane which anchors the principal liposaccarides, lipomannan and lipoarabinomannan through phosphatidyl inositol mannosides. The plasma membrane also interacts with a peptidoglycan layer which is connected to the arabinogalactan through phosphodiester linkages. This assembly forms the core of the mycobacterial cell wall, on which various complex lipids are
tethered. Mycolic acids are esterified to the arabinose sugar to form mycolyl arabinogalactan. As mentioned above, mycolic acids are also acylated to trehalose units to form TMMs and TDMs. Other lipids such as sulfolipids (SL), polyacyl trehaloses (PAT), diacyl trehaloses (DAT), phthiocerol dimycocerosate (PDIM) and mannosyl-β-1-phosphomycoketides interact intensively with mycolyl arabinogalactan and are sometimes shed out of the cell wall [122].

1.8.5.1. SL-1
SL-1 has been implicated in the inhibition of mitochondrial oxidative phosphorylation, alteration of phagosome-lysosome fusion, and stimulation as well as suppression of cytokine and reactive oxygen species production in host leukocytes [120, 122]. However, MTB gene disruption strains lacking fully synthetized SL-1 do not appear to have phenotypes distinguishable from wild type MTB in animal models of infection. In contrast, the diacyl sulfolipid SL1278, a biosynthetic precursor of SL-1, is a well documented active metabolite which was shown to induce an immune response [122].
The sulfotransferase Stf0 initiates SL-1 biosynthesis by sulfating the abundant disaccharide trehalose to form T2S. The acyltransferase PapA2 then catalyzes the esterification of T2S to generate a monoacylated intermediate, SL659. The polyketide synthase Pks2 synthesizes methyl-branched (hydroxy)phthioceranoyl chains using an activated fatty acid starter unit provided by the fatty acid AMP ligase FadD23. PapA1 transfers the product of Pks2 to SL659, yielding diacylated SL1278. Additional acylations of SL1278 are required to produce fully elaborated SL-1. These final steps are chemically similar to the reaction catalyzed by PapA1, but there is no
in vitro evidence that PapA1 is capable of this activity. The lipid transporter MmpL8 has been implicated in SL-1 formation. The MTB ΔmmpL8 gene disruption mutant accumulates the diacyl precursor SL1278 in the cell membrane rather than the predicted SL-1, implying that MmpL8 is required for biosynthesis as well as transport. It was also demonstrated that the final steps in SL-1 biosynthesis and SL-1 transport require Sap (Rv3821) and Chp1 (Rv3822) in addition to MmpL8 [122]. Interestingly, a stf0-deletion mutant was shown to exhibit augmented survival in human but not murine macrophages, suggesting that SL-1 negatively regulates the intracellular growth in species-specific manner [121].

1.8.5.2. PAT/DAT

Mycolipenic acids, the major acyl substituents found in PAT and some forms of DAT, have been shown to be potent inhibitors of leukocyte migration in vitro [123-124]. As well as being B-cells antigens, DAT also inhibit proliferation of murine T cells in vitro. It was also shown that the deficiency of PAT/DAT and PAT affects the surface global composition of mycobacterial cell envelope improving the efficiency with which MTB binds to and enters phagocytic and non-phagocytic host cells [124].

Pks3 and Pks4 are the only enzymes involved in the synthesis of PAT/DAT that have been characterized so far. The targeted disruption of pks3-pks4 resulted in mutants lacking PAT/DAT [123]. PapA3 has been demonstrated to be an acyltransferase required for PAT/DAT biosynthesis [160].

1.8.5.3. PDIM

PDIM were found to be required for the multiplication of MTB during the acute phase of infection [123]. The presence of PDIM in the outermost layers of the envelope was
shown to protect MTB from cidal activity of the reactive nitrogen intermediates produced by activated macrophages. The presence of PDIM also results in down-regulation of the secretion of pro-inflammatory cytokines TNFα by murine macrophages and dendritic cells [123].

Methylated fatty acids of PDIM are synthetized by a multifunctional protein, mycocerosic acid synthase (Mas). The first step in the biosynthesis of phthiocerol involves the transfer of a long-chain fatty acid on protein PpsA by FadD26. Once this starter unit is loaded at the active cysteine residue of the ketosynthase domain and the acyl transferase domain transfers a malonyl group onto the phosphopantetheine arm of the acyl carrier protein domain, a decarboxylative condensation occurs that extends the long-chain fatty acid by a two-carbon unit. This chain is then covalently transferred from the acyl carrier protein of PpsA to the active site KS domain of PpsB through an acyl transfer reaction. The PpsB protein again utilizes another molecule of malonyl-CoA to add a two-carbon unit and generates another hydroxyl group. Similar rounds of chain transfer and extension continue with PpsC and PpsD proteins, which convert the b-keto groups into completely saturated carbons due to the presence of additional dehydratase and enoyl reductase domains. The AT domain of the PpsE protein has broad substrate specificity and can therefore utilize both malonyl- as well as methylmalonyl-CoA extender units, which produces diversity in PDIMs. The assembly and transport of PDIMs require several other proteins. Transesterification of mycocerosic acids with the diol component of phthiocerol is brought about by PapA5. The methoxy group at the 3-position of phthiocerol derives from reduction of a keto group by a stand-alone ketoreductase protein, which is followed by transfer of the methyl group by an Omethyl transferase protein Mtf2. The
mechanism of release of the phthiocerol chain from PpsE is presently unknown. The TesA protein from the Pps cluster is proposed to specifically interact with the PpsE protein to release the covalently acylated phthiocerol chain. The transmembrane protein MmpL7 is proposed to couple the synthesis of PDIM with its export outside the cell envelope, by specifically interacting with PpsE. Recently, it was found that the LppX protein was required for the translocation of PDIM across the plasma membrane [122].

1.9. Project overview
As still not much information is available on the transcriptional regulation of MTB adaptations during latent phase, and on the regulation of changes in lipid metabolism in particular, the goal of this project was to identify the key regulators and regulatory programs responsible for the gene expression in hypoxia/re-aeration conditions and in macrophages. This dissertation describes work performed in support of a larger scale Tuberculosis Systems Biology Project with a focus on the use of ChIP-seq and expression data to reconstruct transcriptional regulation of lipid metabolism in MTB. The project is split into two major aims, each of which is described in one of the subsequent chapters. In Chapter 2 the identification of key regulators and regulatory programs underlying gene expression in MTB during hypoxia and re-aeration is described. Chapter 3 describes the regulatory programs identified in macrophage expression models and their comparison with the regulatory programs in the hypoxia model presented in Chapter 2. The final chapter summarizes the findings described in this work, and suggests potential experiments to follow up on these findings.
Chapter 2. Identification of regulators and regulatory programs underlying gene expression during hypoxia and re-aeration

2.1. Introduction

Although some of the metabolic adaptations of MTB to the host environment are known, the mechanisms underlying these adaptations are poorly understood. The only well-studied regulon involved in the expression changes during hypoxic conditions is DosR. Thus, the first goal of this project was to identify other key regulators and regulatory programs controlling metabolic adaptations to hypoxia, specifically the ones responsible for lipid metabolism.

2.2. Prediction of regulatory networks

The reconstruction of transcriptional networks from high-throughput data can provide valuable information about the regulation of gene expression in cells and cellular processes in general. The number of computational methods that are being developed to reconstruct transcription-regulatory networks (TRNs) is constantly increasing. The main available methods for general TRN inference methods as well as methods for dynamic TRN predictions will be described. Focus will be put on the DREM algorithm as the method chosen for the analysis in this project.

2.2.1. General network inference methods

The large number of possible solutions (the large search space) and the restricted number of independent data points turns TRN and module interference into an undetermined problem. To make the problem of underdetermination more tractable, module and network inference methods apply different strategies to reduce the search space and/or extend the amount of independent information.
2.2.1.1. Module-based versus direct network inference methods

Genes belonging to the same module behave in a similar way under certain environmental conditions. Identification of gene modules involves methods that rely on clustering or biclustering (combining the selection of co-expressed gene sets with a condition selection step to infer the set of conditions that is relevant to the clustered genes). Module-based network inference procedures assign a regulatory program to these modules which drastically lowers the number of interactions to be evaluated during the inference process. Direct network inference methods assign an individual regulatory program to each single gene. A well-known example of a direct method is CLR, an expression-based network inference method that reconstructs an interaction between a TF and a target gene based on the mutual information in their expression [137]. Most module-based approaches not only predict regulatory interactions, but also identify the experimental conditions under which the predicted interactions take place. These methods can assign regulators with expression profiles that are less similar to those of their target genes unlike direct methods. On the other hand, direct methods cover interactions for a larger range of regulators, as module-based inference methods lose interactions with target genes that are not co-expressed with a sufficient number of other target genes [135].

2.2.1.2. Modeling combinatorial regulation

Some methods simplify the problem of modeling combinatorial regulation by assigning regulators to their target genes one by one and composing the combinatorial regulatory programs in a post-processing step that finds sets of regulators which control the same target genes. This stepwise approach makes it impossible to differentiate between truly complex combinatorial regulation, in which
multiple TFs are integrated to trigger the observed gene expression pattern, and condition-dependent regulation, in which different TFs act independently to mediate expression of their target genes under different subsets of conditions. Other methods automatically take into account the condition dependency of the inferred transcriptional programs: regulators that are assigned to the same genes under different subsets of conditions are assumed to act independently, and the ones that are predicted to regulate the same group of target genes in similar conditions, are presumed to act combinatorially, as they are needed simultaneously to trigger the observed co-expression response. Some of these methods can also infer the mode of the combinatorial interactions between the assigned regulators (AND, OR, XOR relationships).

2.2.1.3. Integrative versus expression-based methods

Most non-integrative methods extract the information about regulator-target interaction relying only on expression data itself, assuming that the expression profile of the regulator is an indicator of its activity. This assumption disregards the important role of regulation mechanisms acting on levels other than transcription and the predicted interactions are restricted to cases when regulators are either co-expressed or inversely correlated with their targets. Moreover, most expression-based inference methods cannot distinguish between regulators that actually regulate a gene (i.e. has a direct causal effect) and regulators that are simply co-expressed with a gene. Integrative network inference methods can extend the scope of their predictions beyond interactions that can be inferred from co-expression behavior and usually make more reliable predictions by complementing gene
expression with additional transcriptional information such as motif data or DNA-protein interaction data.

2.2.1.4. Global versus query-driven inference methods

Global module inference methods search for the modules that explain most of the data. This usually involves identifying large pathways that consist of many genes and that are usually responsible for the general responses to major metabolic or condition shifts. Global approaches provide a general view of the active TRN. Query-driven module detection methods, on the other hand, search for genes that are co-expressed, in a condition-dependent way, with a predefined set of genes. These algorithms are deliberately biased towards finding a specific local solution in the search space based on a gene sets chosen specified by the user.

2.2.2. Network prediction methods for time series data sets

The transcriptional response to stresses is a dynamic process involving many TFs that can act at a different time. So just identifying the type of interactions between the TF and the regulated gene doesn’t give us information about the temporal control of the regulatory programs. The availability of time series data sets allows the reconstruction of dynamic regulatory networks which give us valuable information about the behavior of the networks over time and in response to different internal and external stimuli. Several types of methods have been developed specifically for the analysis of temporal data sets.

2.2.2.1. Clustering methods

To obtain a global view of the behavior of genes in time series experiment it is useful to divide (cluster) these genes into smaller groups based on their temporal
expression patterns. Several methods have been developed for clustering of time series data.

One group of methods uses continuous representation of gene expression. Gene expression level can be represented by splines (piecewise polynomials with bounded constraints), piece-wise linear, quadratic, or higher-order interpolation \([135]\). One example is the method proposed by Bar-Joseph et al. \([138]\) which uses B-splines. The clustering method assumes a mixture model where each mixture component corresponds to a cluster, and the expression of each gene is generated through a noisy process from the model expression curve. Bar-Joseph et al. describe a method that simultaneously estimates the parameters for the continuous representation and the assignment of genes to clusters \([138]\).

The second type of methods perform clustering using dynamic features reflecting temporal patterns which are extracted from expression time series. Kim and Kim \([139]\) use first- and second-order differences between adjacent time points as temporal features. Genes are clustered based on the pattern determined by the sequence of features. One limitation of the method is that it requires several replicate experiments and most time series expression data sets are measured with very few or no replicates. Déjean et al. \([140]\) represent genes by smoothing splines, and use the derivatives at some discretization points as features. Genes are clustered by applying hierarchical clustering to the extracted derivatives.

Another group of methods clusters genes using Hidden Markov Models (HMMs). Schliep et al. \([141]\) developed a HMM method to model the dependency between observations of adjacent time points. An HMM is specified by a set of hidden states,
the probability of starting at a given state, the probability of transition from one state to the other, and the probability of generating the gene expression level at each state. The clustering is modeled by a mixture of HMMs, where each HMM corresponds to a cluster. Gene assignment and model parameters are estimated by maximizing the likelihood of the observed expression time series using an Expectation Maximization (EM) style algorithm. The number of clusters is determined by a heuristic procedure that removes clusters with too few genes and splits clusters with too many genes. In [141], a gene is assigned to the cluster corresponding to the most probable HMM. HMM-based methods require that the number of time points to be much larger than the number of states and that is why they work well for long time series but may be problematic to use for short time series.

Clustering of time series data can also be performed by using an autoregressive model [142]. An autoregressive model of order $p$ assumes the expression level at a given time point is a linear function of the expression levels of the same gene in the previous $p$ time points. The clustering algorithm uses an agglomerative procedure to search for the most probable set of clusters. It starts by assuming every expression time series is generated by a different process. In the next step, it computes the model likelihood for all possible pair-wise merges. The method then identifies the merge that results in the highest model likelihood, and, if it is higher than the current model likelihood, merges the two clusters. The procedure stops when the model likelihood cannot be improved by merging anymore.
2.2.2.2. Regression methods for causal inference

Lagged correlations [143] can be used to search for regulatory relationships in time series data. The methods that use this approach can be divided into those that can be applied to a single data set and those that can be applied to multiple data sets at once.

Most of the work applied to a single data set involves regression analysis to identify causal genes. Qian et al. [144] were among the first to use time series data for inferring interactions among genes. To identify causal relationships between a pair of genes they have used local alignment algorithms to find cases where a later expression of one gene matches an earlier expression of another gene and link these two genes. They have also analyzed inverted relationships that could identify repression effects.

Another direction for determining causal relationships from time series data is the use of various graph theory-based methods also known as graphical models. These models include Bayesian networks [145] that have been successfully applied to study static expression data. An extension of Bayesian networks, Dynamic Bayesian Networks (DBNs) can be used to determine regulatory relationships from time series data, often improving on the static version for this type of data [146]. The major challenge associated with learning such networks from data is the condition that a large number of parameters needs to be estimated from a relatively small number of data points (time series experiments are often no longer than 8 time points).

Combining multiple time series data sets when learning DBNs or other time lag models may help in overcoming the dimensionality problem. However, combining
multiple data sets is a non-trivial problem as well. First, sampling rates differ between different data sets, making it hard to determine a common temporal unit for DBNs. Second, for a specific interaction pair (a TF and its target gene) the actual time lag may differ between different experiments since the time scale of the series data may change. Finally, even for a pair of genes displaying time lagged regulation, this relationship might exist in only a subset of the data sets since different pathways may be activated under different conditions. A possible way to combine multiple data sets is to ignore the time lag and rely instead on correlation between the profiles of genes in the data set. This effectively assumes a time lag of 0 for all pairs. For example, Lee et al. [147] used the correlation method to combine a large number of human expression data sets to search for correlated pairs. Another way to address this issue, which is appropriate for combining experiments that study the same system under different conditions (for example, different cell cycle arrest methods) is to align the data sets assuming that genes behave in the same way in all experiments though with different time units. The alignment process determines the appropriate transformation from one time series to another. Once the alignment is determined, the different data sets can be transformed into a common temporal representation and they can then be used to infer DBNs and other lagged models as discussed above. However, when combining more diverse experiments (for example, cell cycle and stress experiments), such an assumption cannot be expected to hold anymore. Thus, DBNs have so far been limited to modeling individual data sets or similar data sets for the same biological system. Shi et al. [148] presented method that may overcome this problem and allow researchers to combine experiments from different conditions in a single DBN. These authors
presented an algorithm that uses a set of known interacting pairs to compute a temporal transformation between every two data sets, regardless of the condition they study. The underlying idea is that some interactions would be present in both data sets and these can be used to learn the temporal transformation between the two data sets. Using an EM algorithm, they align all time series data sets to a common reference data set (usually the longest) and use the aligned experiments to search for additional regulatory interactions, not used in the learning phase, that are present in multiple data sets.

2.2.2.3. Integrative methods

As discussed in the previous section, network inference methods that rely solely on time series gene expression data often face the problem of the number of parameters to fit being much higher than the number of time points. To overcome this problem to a certain extent, inference algorithms can incorporate other data sources to impose additional constraints and reduce the number of feasible models. Adding new types of data to existing models gives rise to its own set of challenges. Such information can be used in a pre- and/or post-processing step to eliminate inconsistent networks or can be tightly coupled with the network inference algorithm, which may require a fundamentally different computational framework. Furthermore, not all types of data are prevalent in certain species. For instance, whereas sequence data are readily available for many species of interest, genome-wide protein–DNA binding studies have only been performed for a few species. In addition, as noted earlier, sequence data are inherently static and protein–DNA binding, protein–protein interactions, and miRNA–mRNA interactions are generally measured at a single time point in a single condition. Thus, it is not always
straightforward to use this information to provide additional insight into dynamic regulatory processes.

Kundaje et al. [149] combined time series gene expression profiles and occurrence counts of known motifs to learn transcriptional modules. Splines were used to model the dynamic expression data, and the modules were learned by using EM to optimize a generative probabilistic graph model. Ramsey et al. [150] extended the time-lagged correlation method to include a motif scanning step. Differentially expressed genes were clustered, a time lagged correlation procedure calculated significance for TF-gene pairs, and the significance scores were combined to yield TF cluster scores. Position-weight matrices were used to scan the promoter regions of the differentially expressed genes and motif enrichments were computed for each cluster. Inferelator [151] first formed biclusters based on gene expression data, regulatory motifs in promoter regions, and a network of functional associations. Kinetic equations were then fit to determine the regulatory impacts between predictor variables, TFs and external stimuli, and the biclusters. This method also models pairwise combinatorial interactions between predictors. An extension to Inferelator [152] adopts a Bayesian approach to improve predictions under long time scales.

While incorporating sequence data is appealing due to its prevalence in many species, motif-based binding predictions are not as informative as experimental protein–DNA binding interaction data. Luscombe et al. [153] presented Statistical Analysis of Network Dynamics (SANDY), a tool for calculating network statistics for dynamic systems. Differentially expressed genes were assigned to a stage in the cell cycle, and an iterative trace-back algorithm was applied to isolate the active TFs and
sub-network at that stage. Sub-networks were subsequently compared based on graph statistics such as topology, presence of network motifs, and TF usage. Lin et al. [154] employed a first-order nonlinear differential equation to combine cell cycle TF binding data and dynamic gene expression data and extract dynamic interactions among the TFs. Several regression-based methods also include protein–DNA binding data to guide the estimation of model parameters. Cokus et al. [155] applied linear regression to time series gene expression data and binding interaction data to estimate dynamic TF activity levels at each time point. The authors then used least squares to estimate a transition matrix that specifies how TFs affect each other’s activity levels over time. Multivariate Random Forests, developed by Xiao and Segal [156], consist of a random forest of multivariate regression trees that use protein–DNA binding and motif data as input and temporal gene expression levels as outcomes. The resulting proximity matrix specifies pairwise gene similarity based on both time series expression and binding information. The authors used the proximity matrix as input to a guided clustering method to identify regulatory cliques.

Probabilistic graphical models have also benefitted from the integration of TF binding information. Dynamic Bayesian Networks, discussed in the previous section, were adapted to include TF binding data as a prior by Bernard and Hartemink [157]. The post-transcriptional modification model presented by Shi et al. [158] learns temporal TF activity levels via a switching model that determines whether a TF is regulated transcriptionally or post-transcriptionally. TFs’ activity levels can then be respectively inferred from either their own gene expression levels or the expression levels of their regulatory targets. Below one type of graphical model applied to this problem, an extension of HMMs, is described in greater detail. This is the method
that was chosen for the prediction of key regulators and regulatory programs during hypoxia and macrophage time courses in this project.

2.2.3. DREM (Dynamic Regulatory Events Miner)

2.2.3.1. The DREM algorithm

The analysis of time series microarray data allows us to observe the dynamics of gene expression and predict the regulatory programs responsible for the behavior of specific groups of genes. DREM is one of the available methods for modeling regulatory networks over time [136]. Based on time-series expression data and TF binding data (ChiP-Seq in our case), DREM infers the global temporal regulatory map – a sequence of paths connecting genes with similar behavior over time. The method identifies so called “bifurcation events” – points on the map where genes with the same expression up to this point diverge, and predicts the TFs controlling these events.

DREM models gene expression as a set of paths where at each time point each path can split into two or more subpaths as a consequence of TF regulation. Genes are associated to a path as a function of the TFs that bind the gene in the regulatory network, and similarity in expression to other genes in the path. The model used by DREM is based on an Input-Output Hidden Markov Model (IOHMM) [136]:
Where $x_t$ is a set of discrete hidden states at time $t$, $u_t$ is an input vector at time $t$, $y_t$ is the output vector at time $t$. The current inputs and the current state distribution are used to estimate the state distribution and the output distribution for the next time step. The input vector for each gene is a vector of transcription factor values, 1 if the TF is predicted to regulate the gene, and 0 otherwise. As each TF is assumed to regulate a gene over the whole time course, the input vector $u_t$ has the same values for each time point $t$, i.e. is static. The probabilistic model $M$ is the following set of parameters: $M = (X, E, \Psi, \Theta, n, \gamma)$.

- $n$ is the number of time points
- $X$ is the set of hidden states. Each hidden state $x_t$ is associated with a Gaussian output distribution, $f_x$.
- $\Theta$ is the set of parameters for the output distributions. For each hidden state $x \in X$ there is an element $(\mu_x, \sigma_x) \in \Theta$ corresponding to the mean and standard deviation of the Gaussian distribution, $f_x$.

\[
\begin{align*}
    x_t & = f(x_{t-1}, u_t) \\
    y_t & = g(x_t, u_t)
\end{align*}
\]
- $E$ is the set of directed edges connecting hidden states of $X$, each edge is a valid transition among hidden states. This set of edges is constrained to enforce a tree structure. Each hidden state can have at most $\gamma$ children. Formally the root of the tree is assumed to have only one state at time point 0.

- $\Psi$ contains the parameters controlling transition probabilities between hidden states. If a state $x \in X$ has 2 or more children, then there is an element $\psi_x \in \Psi$ where $\psi_x$ is a vector of parameters for a logistic regression classifier.

Let $o_g = (o_g(1), \ldots, o_g(n-1))$ be the log ratio expression values for gene $g$ at time points 1 to $n-1$ relative to a time point 0 control. The transition probability for a gene $g$ to transition state $x_b$ at time point $t$ given that it is in hidden state $x_a$ at time point $t-1$ is $P(X_t = x_b | X_{t-1} = x_a, u_g)$. The logistic regression classifier associated with state $x$ maps the static input vector $u_g$, to the transition probabilities for a hidden state $x_a$.

The likelihood density, $L$, for a set of genes $G$ for the model $M$ is

$$L(G|M) = \sum_{g \in G} \log \sum_{q \in Q} \prod_{t=1}^{n-1} f_{q(t)}(o_g(t)) \prod_{t=1}^{n-1} P(H_t = q(t)|H_{t-1} = q(t-1), u_g)$$

$Q$ is the set of all paths of hidden states of length $n$ starting from the root with a non-zero probability. For a path $q \in Q$, $q(i)$ is the hidden state of the path at time point $i$.

The parameter learning step in the model is performed by using a modification of the Baum-Welch training algorithm for a regular Hidden Markov Model. During each maximization step of the Baum-Welch algorithm each logistic classifier is retrained.

When training a classifier, for every gene in the training set the classifier is given the
probability that the gene is going through the corresponding hidden state based on
the current values of all the parameters in the model.

To find the optimal structure for the model, \((X, E)\), a search starts from a single chain
of hidden states followed by a search over various structures. To train and test these
structures, the algorithm splits the set of genes into two sets, \(G_{\text{train}}\) and \(G_{\text{test}}\).
\(G_{\text{train}}\) contains 75% of the genes. The remaining 25% of genes, \(G_{\text{test}}\), are used for
scoring the structures. During the parameter maximization the search considers
adding and deleting paths to the model while \(L(G_{\text{test}} \mid M)\) increases. After no more
paths can be added or deleted from the model without improving the test score, the
weakly supported paths that might be overfitting \(G_{\text{test}}\) are removed through
randomly resplitting the set of genes used for training and testing. The retrained
model with the path deleted is compared to the current one. If the score for the new
model does not change significantly, the old model is replaced by the new one. A
similar procedure is used to delay splits for which the split could be placed one time
point later with no significant effect on the score.

The algorithm then combines the train and test sets and retrains the model on the
combined set without changing the structure. Genes are assigned to their most likely
paths using Viterbi algorithm. Paths which have less than 5 genes assigned are
removed from the model. The splits and paths are then scored for association with
TFs. The score for a TF is computed using the hypergeometric distribution:

\[
\min_{i=c_A} \sum_{i=c_A}^{\min(c_S,n_A)} \left( \frac{c_S}{i} \frac{n_A-c_S}{n_A-i} \right) \frac{n_S}{n_A}
\]
where $S$ is the split, $A$ is a path out of the split, $n_S$ is the total number of genes assigned to the path into the split, $n_A$ is the total number of genes on path $A$ out of the split, $c_S$ is the total number of genes into the split predicted to be regulated by a TF, $c_A$ is the number of genes on path $A$.

2.2.3.2. Advantages and limitations of DREM

Biological systems are dynamic. The ability of DREM to derive dynamic maps that associate TFs with the genes they regulate and their activation time points may lead to a better understanding of the system studied and potential prediction of new regulatory roles for some of the TFs. As DREM infers the time points at which TFs regulate their targets, it can differentiate master regulators that control the immediate response to a stimulus from secondary regulators that are active later. Through determining the time at which a TF is most active DREM can also identify the best time to conduct binding experiments for particular TFs to experimentally verify their role in the stimulus response.

DREM is highly dependent on the input data. The sampling rate may play an important role in the ability of the algorithm to derive accurate regulatory maps. The models identified by DREM are currently limited to tree structures. DREM also does not explicitly model regulation through other mechanisms, such as chromosome remodeling and mRNA degradation. Transition probabilities are computed using logistic regression, which does not capture all types of combinatorial interactions. Another limitation of DREM is that the output dynamic path model is sometimes chosen from a number of possible dynamic maps with similar scores.
2.3. Data sets

Hypoxia and re-aeration time course expression data for Nimblegen was generated by our collaborators (protocol described in [63], GEO accession number GSE43466) and used in the analysis. The whole time course expands from day 0 to day 14. MTB is under hypoxia from day 0 to day 7 and under re-aeration from day 7 to day 14 (Fig.3).

![Transcriptomics](image)

**Figure 3. Hypoxia profiling.** The hypoxia time course is colored in blue, the re-aeration time course – in yellow. The red arrows show the time points where samples were taken.

TF binding data from ChIP-Seq experiments for over 80 TFs was used as the second type of input for the DREM algorithm. To define the regulatory role of a TF an additional overexpression data set of over 200 TFs was used.

2.4. Data processing and statistical methods

2.4.1. Microarray data processing

The hypoxia and re-aeration time course expression data for Nimblegen is a combination of 3 separate experiments – 2, 6, 7. The number of replicates used per
experiment is shown in the table below. RMA (Robust Multichip Average) with background correction and quantile normalization was applied over all probes together. The expression values for each day in hypoxia/re-aeration were obtained by averaging over the RMA values of all available replicates from corresponding experiments. The fold changes are log2 ratios of each day to day 0. P-values for differential expression were adjusted with Benjamini-Hochberg method.

<table>
<thead>
<tr>
<th>Day</th>
<th>Total # of replicates</th>
<th>SG2 replicates</th>
<th>SG6 replicates</th>
<th>SG7 replicates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day 0</td>
<td>11</td>
<td>3</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Day 1</td>
<td>11</td>
<td>3</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Day 2</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Day 3</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Day 5</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Day 7</td>
<td>11</td>
<td>3</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Day 8</td>
<td>11</td>
<td>3</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Day 9</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Day12</td>
<td>4</td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>Day14</td>
<td>4</td>
<td></td>
<td></td>
<td>4</td>
</tr>
</tbody>
</table>

Table 1. Number of replicates per experiment in the hypoxia/re-aeration expression data set.

2.4.2. TF binding list

The TF binding list was generated based on ChIP-Seq data under the threshold of 1% of maximum impulse coverage for a TF. TF binding sites in intergenic upstream, genic upstream and genic in-gene regions were considered. After decomposing the
time course expression data into a set of expression paths we compared the path, the expression of each TF binding genes in the path, and the predicted regulatory role of the TF from the regulatory network to assess the degree to which expression patterns might reflect the direct action of transcription factors. To define the regulatory role of the TF in the network (activator or repressor) Z-scores from the overexpression data set were used. The TF is considered to be an activator if Z-score > 1, a repressor – if Z-score < -1. If Z-score Є [-1, 1], that is considered as a weak binding event. Z-score values are generated over the distribution of genes whose behavior was affected by the overexpression of the TF.

2.4.3. DREM parameters

For building regulatory expression models in this and next chapter the main DREM parameters were used with the following values: minimal absolute expression change – 1, maximal number of paths out of a split – 3 (higher numbers would generate splits that lead to paths with very few genes), convergence likelihood – 0.01%. Predictions of TFs regulators of bifurcation points with p-values < 0.05 were considered significant.

2.5. Regulatory programs predicted during hypoxia

DREM was applied to decompose the time course expression data in a nested set of paths (Fig. 4, DREM results in Appendix A, Table 2: A). Most of the genes during hypoxia are down-regulated. Still, we see 3 paths in which genes are up-regulated (especially path1-1). These paths have a considerable number of TFs, and are enriched in cholesterol metabolism, sulphate reduction and methylcitrate cycle. The
paths of down-regulated genes are enriched in respiration, glycolysis, pentose phosphate pathway, cell wall and amino acid biosynthesis.

Splits (bifurcation points) in DREM expression paths are observed in the first three days of the hypoxia time course. DosR and KstR are predicted to be responsible for the separation of gene expression patterns in during all three days (Fig. 5). Rv0081, Rv1990c, Rv0324, Rv0967, WhiB1 are involved in the change of gene expression during day 1 and 2, FurB – day 2 and day 3, the three-component system regulator TcrA – day 2, the two-component system regulator MtrA – day 3.

Figure 4. DREM paths during hypoxia/re-aeration. Each path (average expression of all the genes in the path over the time course) is assigned a number based on the order and amount of preceding bifurcation points. The hypoxia time course is colored in cyan, the re-aeration time course – in yellow. Pathway enrichment categories in the upper right corner refer to upregulated paths, the ones in the bottom right corner – to downregulated paths.
Figure 5. TFs predicted by DREM to be responsible for separation in gene expression patterns during the hypoxia time course. The days when a bifurcation in expression occurs are named with subsequent numbers.
2.6. Validation of key regulators and regulatory programs predicted by DREM

Overexpression microarrays were used as an independent data set to validate some of the key regulators and regulatory programs predicted by DREM (by incorporating TF binding data and time course expression data). We assessed the degree to which expression patterns might reflect the direct action of transcription factors by evaluating the consistency between the path, the expression of each TF binding genes in the path, and the predicted regulatory role of the TF from the regulatory network (based on overexpression data).

Using this approach, we correctly identify DosR an activator of path 1-3 (Fig. 6). This path corresponds almost entirely to the DosR regulon [17] (43 out of 48 genes in the regulon are in the path), and we correctly predict an activator role for DosR and see positive correlation between DosR and the genes in this path.

![Figure 6. DREM path 1-3 regulated by DosR.](image)

The path is colored in dark blue, the expression of DosR – in orange, the expression of the genes in the path which are bound by DosR – in light blue. The Venn diagram shows how many of the genes assigned to path 1-3 belong to the DosR regulon.
For KstR, a repressor with known regulon [72], we identify two types of paths that KstR is predicted to be the regulator of: a downregulated and an upregulated one (Fig. 7, A-B). KstR is upregulated during the hypoxic time course and the genes bound by the TF in path 3-1-2 are downregulated (Fig. 7, A). Thus, in this case we predict correctly KstR as a repressor of the genes in this path. Path 1-1 has two genes with high positive Z-scores for KstR and several genes with low positive Z-scores (weak binding), i.e. we predict KstR to be also an activator of some genes in this path. However, there is also a group of genes that belong to the KstR regulon and are known to be repressed by the TF, and still they are upregulated. One reason for that can be combinatorial binding, as many of these genes are bound by Rv0081 and Rv0023 which are activators of these genes. Another hypothesis is that there is a factor that leads to the de-repression of KstR. Cholesterol is known to de-repress KstR but because no cholesterol was present in the media during the hypoxic time course, a potential source of de-repression can be other fatty acids produced in the media, such as propionate.

The third TF for which DREM regulatory predictions were validated, is FurB – a Zn-dependent repressor with known regulon [162]. FurB was predicted by DREM as one of the regulators of the most downregulated path 3-2 (Fig. 8). The genes bound by FurB in this path all belong to the regulon of FurB (9 out of 34 genes) and FurB is upregulated during the hypoxic time course (Fig. 8). Thus, FurB was validated as a repressor of the genes it binds in path 3-2.
Figure 7. DREM paths regulated by KstR. A. Downregulated path 3-1-2. B. Upregulated path 1-1. Paths are colored in dark lila (A) and pink (B), the expression of KstR – in green, the expression of the genes in the path bound by KstR: high positive Z-scores – red, low negative Z-scores – dark blue, weak binding – gray.
Figure 8. DREM path 3-2 regulated by FurB. The path is colored in dark blue, the expression of FurB – in orange, the expression of the genes in the path which are bound by FurB – in light blue.

2.7. Identification of Rv0081 as a candidate high-level regulator broadly predictive of the overall expression of a set of genes during hypoxia and re-aeration

Strikingly, we also identify Rv0081 as a candidate high level regulator broadly predictive of the overall expression of sets of genes during hypoxia and re-aeration (Fig. 9). In particular, the regulatory role of Rv0081 with respect to individual genes, as determined independently by induction, matches the correlation in expression between Rv0081 and these genes during hypoxia and re-aeration. Rv0081 is initially induced during hypoxia, declines in expression throughout hypoxia, and is expressed at low level during re-aeration. Rv0081 binds 23% to 41% of genes shown in Figure 9. For path 1-1, which displays genes that are highly activated during hypoxia and trend down during re-aeration, 70% of Rv0081 binding is predicted to result in activation. Conversely, for path
Figure 9. Rv0081 activation/repression ratios for each DREM path during hypoxia. Each path is colored according to the color scheme in Fig. 4. The name of each bar consists of the name of the path index and the ratio of the number of genes bound by Rv0081 in the path to the total number of genes in the path.

3-2, which is highly repressed during hypoxia, 85% of Rv0081 binding is predicted to be repressive. Paths with intermediate levels of activation or repression display mixes of activating and repressing Rv0081 binding (Fig. 9). As a whole, almost a linear positive correlation is observed between the number of genes activated/repressed by Rv0081 and the behavior of all the genes in the DREM path which suggests that Rv0081 regulatory roles are predictive of the expression of genes in each path. Thus, a broad regulatory role of Rv0081 is supported by two independent sources of evidence: overexpression of RV0081 alters the expression of
a large number of genes and the regulatory role of Rv0081, predicted based on ChIP-Seq that reveals a large number of binding sites (which are also detected when ChIP is performed on Rv0081 during hypoxia) and the time course expression, correlates with the expression of bound genes in an independent expression data set derived from a significantly different condition.

Within each path, however, there is substantial variance in expression and not all genes are bound by Rv0081. Part of the variance can be explained by the pattern of binding of other TFs. In most of the paths we can see that a considerable number of the genes are bound by several transcription factors.
Chapter 3. Comparison of regulatory programs during hypoxia and macrophage infection

3.1. Introduction

Adaptations to hypoxia has been shown to play an important role in MTB survival in the host. However, the pathogen faces other stresses in the phagosome like nitrosative stress and pH change. So far in terms of regulation it has been only shown that the DosR regulon is induced by NO as well [129-130]. Thus, the second aim of this project was to generate a macrophage expression model and compare it with the hypoxia model to check how different the expression of genes is in the two models and what key regulators are responsible for the common or different regulatory programs controlling these genes.

3.2. Profiling of MTB under conditions other than hypoxia

3.2.1. Transcriptional response in macrophages

MTB has been shown to adapt to the phagosome environment of naïve and activated macrophages by the induction of fatty acid-degrading enzymes and DNA repair proteins, the remodeling of its envelope, and the production of secreted siderophores to facilitate the acquisition of iron [66, 124-128]. Schnappinger et al. [66] reported the expression of MTB genes in murine macrophages over short-term infection out to 2 days. Several functional gene groups were identified that are responsible for the MTB metabolic adaptations. The mbt operon, which directs the synthesis of mycobacterial siderophores, was upregulated both in resting and activated macrophages. Most of these genes were also induced in vitro by NO and H₂O₂, supposedly due to oxidative damage to IdeR, a repressor of the MTB iron
scavenging response. Thus, the increased induction of low iron response genes in activated macrophages might be caused by NO rather than a decrease in the availability of iron in this environment. If that is the case, the NO-mediated induction of the iron uptake system could aid in the repair or replacement of dam.ages iron-containing proteins as a result of the exposure to oxidative and nitrosative stress [66]. The ability of MTB to survive in the macrophage phagosome suggests that remodeling of the mycobacterial surface should take place. FadD26 and UmaA, which might be involved in such remodeling, were identified within the differentially expressed genes in the intraphagosomal transcriptome. FadD26 is involved in the synthesis of PDIM, which, as mentioned above, serves as a mycobacterial virulence factor. UmaA, a protein of unknown function, belongs to a highly conserved family of mycolic acid methyl transferases that modify lipids of the mycobacterial cell wall and might be important for MTB virulence [66]. Remodeling of the cell envelope appears to be accompanied by a DNA recombination and repair response because alkA, recX, recG, dinF, and radA were shown to be induced in the phagosome. Moreover, the macrophage appears to induce DNA damage independently of NO, as these genes were similarly regulated in wildtype and NOS2-deficient macrophages.

To survive and replicate in naive macrophages MTB must assimilate carbon and produce energy in the phagosome. The fact that MTB switches its carbon source from glucose and glycerol during in vitro growth to fatty acids in the phagosome was supported by the induction of A (CoA) synthase (fadD3,9,10,19), acyl-CoA dehydrogenase (fadE5,14,22–24,27–29,31), enoyl-CoA hydratase, (echA19), hydroxybutyryl-CoA dehydrogenase (fadB2,3), and acetyl-CoA transferase (fadA5,6) [66]. The induction of multiple genes for four of the five transformations suggests that
different isoenzymes are required to catabolize structurally diverse fatty acids. Breakdown products of fatty acids (acetyl-CoA and propionyl-CoA) are likely metabolized via the citric acid and glyoxylate cycles as suggested by the induction of the genes \( gltA1 \), Rv1130, and \( icl \). Induction of \( pckA \), which encodes the gluconeogenetic rate-limiting enzyme phosphoenolpyruvate carboxykinase, suggests that fatty acids are in part converted into sugars via gluconeogenesis. \( FadE5 \), \( echA19 \), \( icl \), and \( pckA \) and the protein product of Rv1130 were shown to be induced in human macrophages or macrophage-like cell lines as well [134].

Changes in energy metabolism during macrophage infection are accompanied by changes in intermediary metabolism. Genes that encode subunits of NADH dehydrogenase (NDH)-1 (\( nuoABDFIL \)), the ubiquinol-cytochrome C complex (\( qcrCA \)), and ATP synthase (\( atpABDEFGH \)) were shown to be repressed in both murine and human macrophages [66, 124-128]. The repression of genes encoding NDH-1, the cytochrome C complex, and the ATP synthase presumably reflects the reduced need for energy generation during bacteriostasis, the growth state of intraphagosomal MTB, compared with logarithmic growth. By contrast, \( frdA \), \( narX \), and \( ndh \), encoding fumarate and nitrate reductases and NDH-2, respectively, were shown to be induced. The modulation of the respiratory chain that likely occurs as a consequence of the induction of \( frdA \), \( narX \), and \( ndh \) could reduce the bioenergetic efficiency of the respiratory chain. Alternatively, the induction of these genes could signal an increased need for NAD regeneration, the second major function of the respiratory chain besides energy production. L-oxidation of fatty acids consumes NAD and FADH in every cycle of fatty acid oxidation. Induction of NDH-2, thus, might represent an adaptation necessary to ensure NAD levels that are sufficient to allow
increased L-oxidation of fatty acids during bacteriostasis. The induction of *ald*, which encodes an l-alanine dehydrogenase, might also contribute to the regeneration of NAD because the reductive amination of pyruvate is the kinetically preferred reaction at physiological pH [66].

More recently, Rohde et al. [126] showed the induction of the same genes involved in carbon and fatty acid metabolism, respiration during long-term infection of murine macrophages. The group also detected the induction of 17 members of the PhoPR regulon, including the acid responsive *aprABC* locus. The activation of ~60% of the DosR regulon and the majority of KstR regulon, including *mce4* and *igr* loci, was shown as well. While PDIM and mycolic acids appeared to be downregulated during the time course, *pks2-papA1* (SL-1) and *pks3-pks4-papA3-mmpL10* (PAT/DAT) were induced. The enhanced synthesis of SL-1 and PAT/DAT may serve as a sink to prevent accumulation of the toxic propionyl-CoA, a byproduct of intracellular fatty acid and cholesterol metabolism. *WhiB3* as well as *espABC* (a locus required for ESX-1) and its regulator *espR* were also shown to be induced during the time course [126].

Another group screened for transposon mutants survival in murine macrophages and identified essential genes involved in fatty acid metabolism, transporters, components of ESX-1 secretion apparatus [126].

In human THP-1 macrophage cell lines many of the genes involved in carbon, fatty acid and iron metabolism, respiration and heat shock overlap [133]. Additional groups of genes that were identified: (1) upregulated: *mymA* operon (restructuring of cell envelope), *cysD* and *cysN* (synthesis of 3'-phosphoadenosie-5' incorporated into

3.2.2. Transcriptional response to reactive nitrogen and oxygen species
The DosR regulon was shown to be induced under low concentrations of NO in murine macrophages [129-130]. The induction of these genes by both hypoxia and NO suggests that MTB encounters one or both of these stresses during the infection course in the macrophage. A group of genes that are responsible for oxidative stress defense were induced by high concentrations of NO as well as by H$_2$O$_2$ [131]. This group includes \textit{katG}, \textit{ahpC}, \textit{trxC}, \textit{trxB2}, \textit{sigB} and \textit{sigH}. Both NO and H$_2$O$_2$ also were shown to strongly induce genes involved in iron acquisition (\textit{mbt} cluster) while \textit{bfrB} gene, encoding the iron-storage protein bacterioferrin, was repressed [130]. The genes responding specifically to H$_2$O$_2$ include: \textit{recA}, \textit{radA}, \textit{alkA}, Rv3201c, Rv3202c (DNA repair), several PPE/PE genes (PE20, PPE31, PPE32, PPE33, PE33), Rv1130 ad \textit{gltA1} (propionate metabolism), \textit{leuC} and \textit{leuD} (leucine biosynthesis). 12 genes were shown to be induced specifically under NO exposure including \textit{sigE}, \textit{cyp138}, Rv0967, \textit{lpqA}.

3.2.3. Transcriptional response to nutrient starvation
The following functional groups of genes were shown to be differentially expressed in a nutrient starvation model [131]: (1) upregulated: \textit{pdhABC} (pyruvate
dehydrogenase), frdA (fumarate reductase), pks10, pks2 (SL-1 synthesis), sigB, sigE, sigF and sigD (sigma factors), Rv2034, Rv1152, Rv3291c, whiB3, Rv2358 (TFs), kdpE (a two-component system), subl-cysT-cysW-cysA (sulphur metabolism), pirG (cell envelope); (2) downregulated: gap and tpi (glycolisis), fum, can and icd1 (TCA cycle), nuoA-M, atpA-H (respiration), fas, desA2, desA3, cmaA1 and umaA2 (mycolic acid biosynthesis), ppsABCDE and mas (PDIM biosynthesis).

3.2.4. Transcriptional response to acidic conditions
Under acidic conditions (low pH) in vitro [132], the following main groups of genes were shown to be differentially expressed: (1) upregulated: Rv3083-Rv3089 operon (peptide synthesis), R1130, Rv1131, fadB2, icl (fatty acid metabolism); (2) repressed: Rv3249c (TF), fabD, acpM, kasA, kasB and accD6 (extension of mycolic acids).

3.2. Data sets
Initially, four different data sets of MTB during macrophage infection were used: three microarray data sets (Schnappinger et al. [66], Russell et al. [126] and Tailleux et al. [127]) and one RTPCR data set generated by our collaborators. All of these data sets are time course experiments, where the time points range from hours to days. As the macrophage time course generated by the Russell’s group is on the same time scale and has the highest overlap with the hypoxia/re-aeration time course in terms of time points (Fig. 10), we chose this data set for the comparison of the hypoxia and macrophage models obtained with DREM. Additional microarray data set representing MTB under NO and H$_2$O$_2$ stress [130] was used for comparison with the hypoxia and macrophage data sets.
3.3. Data processing and statistical methods

3.3.1. Microarray data processing

For the microarrays for which raw data was available background correction and loess normalization were performed, the expression values of each gene were averaged over replicates and fitted to a linear model in R’s limma package. For the rest of the microarrays the log ratios provided by the corresponding group were used. In all data sets the final log ratios are a result of the comparison of the expression values of each time point to the expression values of MTB grown in log phase in normal conditions.

3.3.2. Comparison between expression data sets

As the microarrays to be compared with each other are from different platform (the hypoxia data set is Nimblegen, the other arrays are two-color) and experiments, they can’t be compared directly with each other. For the comparison of the log-fold changes from each microarray, the values of microarrays with significantly different mean and standard deviation from the rest of the data sets, were scaled accordingly using the formula:
\[ X'(k) = X(k) \times \frac{\sum_{i=1}^{n} \text{std}(i)}{\text{std}(k)} \]

Where \( k \) is the index of the \( k \)-th dataset to be scaled, \( \text{std}(i) \) – standard deviation of expression values in \( i \)-th microarray, \( X(k) \) – an array of expression values \( k \)-th microarray.

3.4. Pathway enrichment of hypoxia and macrophage models obtained with DREM

The expression paths in the hypoxia and macrophage models (DREM results in Appendix A, Table 2: A-B) were divided into upregulated and downregulated based on their gene behavior. First, each of these paths was analyzed for metabolic pathway functional enrichment to see if there are paths in the hypoxia and macrophage models which are enriched for the same functional categories suggesting a similar mechanism for the corresponding metabolic processes. Second, we checked if any of the enriched paths have the same regulator assigned to them by DREM.

Both hypoxia and the macrophage model are enriched in methylcitrate cycle, cholesterol catabolism (upregulated paths) and respiration (downregulated paths). Methylcitrate cycle is one of the toxic propionyl-CoA assimilation which explains why upregulated paths in both systems are enriched in genes involved in the cycle. The modulation of the respiratory chain caused by the switch from aerobic to anaerobic conditions, leads to the downregulation of genes responsible for aerobic respiration in both hypoxia and macrophage models.

The two upregulated paths in both hypoxia and macrophage model, which are enriched in cholesterol catabolism, were also predicted by DREM to be regulated by the same TF – KstR. Although KstR, which is a repressor, is induced during hypoxia
and in the macrophage, and there is no cholesterol in the media in the experiments, de-repression of cholesterol degradation genes is observed. As mentioned in chapter 2, one hypothesis is that short or intermediate chain length fatty acids might be generated during hypoxia and in the macrophage, resulting in the de-repression of KstR and the induced expression of cholesterol degradation genes. Additionally, as a mechanism, although cholesterol is not present in the media, the cholesterol degradation pathway may be activated as a result of being coupled with the redox stress pathway. Rv0081 which is one of the regulators in the redox stress pathway (together with DosR) binds many of the genes (as an activator) in these two paths in the hypoxia and the macrophage models.

Upregulated paths in macrophage models are additionally enriched in genes involved in the response to nitrosative stress and lipid metabolism. This observation can be explained by the fact that in the macrophage nitrosative stress is present unlike in hypoxia, and some genes responsible for complex lipid synthesis (PAT/DAT and SL-1, results presented in 3.6) are upregulated in the macrophage models but downregulated in hypoxia.

3.5. Regulators predicted to be responsible for the expression of genes in the early stages of both hypoxia and macrophage models

Both in hypoxia and macrophage models DosR and KstR were predicted as the regulators responsible for the expression of genes in the early stages (up to day 2) of the time courses (Fig. 11). This prediction agrees with the finding of DosR being the regulator of the initial hypoxic response [61, 64] (and also induced by NO [65]). KstR, which controls genes responsible for cholesterol degradation and lipid metabolism
and is induced in both hypoxia and the macrophage, is potentially involved in the initial stage of the response in both types of models as well.

Figure 11. DREM predictions for DosR and KstR as key regulators during the early stages of both hypoxia and macrophage models. KstR and DosR were predicted as key regulators up to day 2 in both hypoxia and macrophage time courses. The color of the boxes corresponds to the color of the path the TFs are assigned to in the model.

3.6. Genes, involved in the synthesis of SL-1, PAT/DAT and regulated by PhoP, have opposite expression patterns in hypoxia and macrophage models

A group of genes that are involved in the synthesis of complex lipids: pks2, papA1 (SL-1 synthesis), pks3, pks4 and papA3 (PAT/DAT synthesis), was identified to be downregulated in the hypoxia model and upregulated (differentially expressed) in the macrophage model (Fig. 12). These genes belong to the two most upregulated DREM paths in the macrophage model (Fig. 13). Moreover, PhoP is predicted by DREM as the regulator of one of these paths (the most upregulated one). All of these genes are also downregulated in the PhoP mutant [78] and pks2, pks3, pk4 and papA3 have been reported to be part of its regulon [89] and are bound by PhoP in our Chip-Seq data. PhoP stay at low levels in hypoxia but is upregulated in the macrophage (Fig. 14, A).
Figure 12. Clustergram of the group of genes which has a different behavior in hypoxia and macrophage models. Hypoxia/Re-aeration – the hypoxia/re-aeration time course, Macrophage – macrophage infection time course expression data generated by Russell’s group, PhoP KO – PhoP knock-out microarray, NO – expression data set where MTB was exposed to different concentrations of nitric oxide (nitrosative stress).

Figure 13. Genes involved in complex lipid biosynthesis and the upregulated DREM paths in the macrophage model that they are associated with.

Interestingly, this group of genes is downregulated or does not change significantly in the microarray data set where MTB was exposed to nitrosative stress. Additionally, the action of PhoP has been shown to be pH-dependent [78]. Thus, one hypothesis is that PhoP is the potential regulator of the genes involved in SL-1 and PAT/DAT.
biosynthesis as a result of pH change that takes place in the macrophage environment but not during hypoxia. It’s possible that PAT/DAT and SL-1 precursors (the transporters \textit{mmpL8} and \textit{mmpL10}, which are needed for the final step of the lipid synthesis, are not upregulated in the macrophage models) serve only as a metabolic adaptation (reserve energy and carbon source) to the granuloma environment. In murine models of chronic and progressive pulmonary TB SL-1 and PAT/DAT have been shown to be over-produced possibly because they play a role in virulence [161].

3.7. Additional regulators whose action can explain the specific behavior of genes responsible for SL-1, PAT/DAT synthesis in hypoxia versus macrophage infection

Based on our TF network, PhoP is only one of the TFs which bind the genes responsible for complex methyl-branched lipid biosynthesis. A more complicated mechanism may be involved in the regulation of \textit{pks2} and \textit{pks3} polyketide synthases. In our binding network PhoP binds WhiB3 and both TFs bind \textit{pks2} and \textit{pks3} and have been shown to modulate the production of PAT/DAT and SL-1 [78, 89, 92, 93]. In the hypoxia model PhoP stays at low levels while WhiB3 is upregulated, but both TFs are upregulated in the macrophage models (Fig. 11) and that is where upregulation of \textit{pks2} and \textit{pks3} is observed. That observation leads to the hypothesis that PhoP and WhiB3 form a feed-forward loop (FFL) and both TFs need to be activated for the induction of \textit{pks2} and \textit{pks3} which control the first step in the biosynthesis of SL-1 and PAT/DAT respectively.
Figure 14. PhoP and WhiB3 time course expression and FFL. A: Expression in hypoxia and RTPCR macrophage models. B: FFL structure.
Chapter 4. Discussion

4.1. Conclusions

This work describes the identification of temporal expression trends during hypoxic and macrophage time courses and their association with possible key regulators and regulatory programs with the incorporation of microarray, ChIP-Seq and TF overexpression data. Thus, the first dynamic regulatory models for hypoxia and macrophage infection were built. The results presented confirm already known key regulators like DosR, KstR and FurB (predictions were validated on the overexpression data set) as well as provide new insight about regulators like Rv0081 for which no data has been reported in the literature. Rv0081 was found to be a candidate high level regulator broadly predictive of the overall expression of sets of genes during both normoxia and hypoxia. DosR and KstR were predicted as the regulators responsible for the expression of genes in the early stages (up to day 2) of both hypoxia and macrophage models. Along with similar trends in the expression of genes involved in respiration, cholesterol catabolism and methylcitrate cycle, a direct comparison between the hypoxia and macrophage expression models uncovered a potential condition-specific regulatory program controlling a group of genes responsible for the synthesis of complex methyl-branched lipids (PAT/DAT and SL-1). PhoP is the potential regulator in this program predicted to activate these genes as a result of pH change which takes place in the macrophage environment but not during hypoxia. A more complex mechanism for pks2 and pks3 might involve the cross-talk between PhoP and WhiB3 forming a FFL identified in our transcriptional regulatory network.
4.2. Future directions

As one of the central findings in this work is connected to revealing the transcriptional regulatory mechanisms behind metabolic adaptations of MTB involving complex lipids biosynthesis, the hypothesis of PhoP being a condition-specific regulator of SL-1 and PAT/DAT synthesis can be followed up by experimental validation. One of the experiments for the validation of this hypothesis can be profiling wild type MTB with and a PhoP knock-out at least in two conditions – hypoxia and acidic pH, with additional conditions being NO exposure and macrophage infection. Chromatography experiments with complex lipid extraction (SL-1 and PAT/DAT) in the same conditions can be conducted as well.

A method like single cell microfluidics can be used to identify and analyze the dynamics of the potential FFL between PhoP and WhiB3.
### APPENDIX A

**Table 2. DREM TF predictions for hypoxia and macrophage time courses.**

<table>
<thead>
<tr>
<th>Path Name</th>
<th>TF</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>path1</td>
<td>Rv3133c</td>
<td>1.45E-18</td>
</tr>
<tr>
<td>path1</td>
<td>Rv1255c</td>
<td>1.45E-05</td>
</tr>
<tr>
<td>path1</td>
<td>Rv0324</td>
<td>8.11E-04</td>
</tr>
<tr>
<td>path1</td>
<td>Rv1994c</td>
<td>0.01</td>
</tr>
<tr>
<td>path1</td>
<td>Rv0474</td>
<td>0.014</td>
</tr>
<tr>
<td>path1</td>
<td>Rv1985c</td>
<td>0.023</td>
</tr>
<tr>
<td>path1</td>
<td>Rv3574</td>
<td>0.032</td>
</tr>
<tr>
<td>path1</td>
<td>Rv1956</td>
<td>0.035</td>
</tr>
<tr>
<td>path1</td>
<td>Rv2610c</td>
<td>0.039</td>
</tr>
<tr>
<td>path1</td>
<td>Rv0081</td>
<td>0.046</td>
</tr>
<tr>
<td>path2</td>
<td>Rv2021c</td>
<td>2.91E-03</td>
</tr>
<tr>
<td>path2</td>
<td>Rv0465c</td>
<td>4.58E-03</td>
</tr>
<tr>
<td>path2</td>
<td>Rv1990c</td>
<td>0.041</td>
</tr>
<tr>
<td>path3</td>
<td>Rv0967</td>
<td>4.31E-09</td>
</tr>
<tr>
<td>path3</td>
<td>Rv3249c</td>
<td>7.18E-07</td>
</tr>
<tr>
<td>path3</td>
<td>Rv1049</td>
<td>1.47E-05</td>
</tr>
<tr>
<td>path3</td>
<td>Rv3219</td>
<td>2.99E-03</td>
</tr>
<tr>
<td>path3</td>
<td>Rv0767c</td>
<td>3.00E-03</td>
</tr>
<tr>
<td>path3</td>
<td>Rv2324</td>
<td>4.35E-03</td>
</tr>
<tr>
<td>path3</td>
<td>Rv2887</td>
<td>9.95E-03</td>
</tr>
<tr>
<td>path3</td>
<td>Rv2034</td>
<td>0.027</td>
</tr>
<tr>
<td>path3</td>
<td>Rv0653c</td>
<td>0.042</td>
</tr>
<tr>
<td>path1-1</td>
<td>Rv0602c</td>
<td>3.50E-03</td>
</tr>
<tr>
<td>path1-1</td>
<td>Rv2324</td>
<td>0.023</td>
</tr>
<tr>
<td>path1-2</td>
<td>Rv3574</td>
<td>1.16E-03</td>
</tr>
<tr>
<td>path1-2</td>
<td>Rv1985c</td>
<td>7.81E-03</td>
</tr>
<tr>
<td>path1-2</td>
<td>Rv1990c</td>
<td>0.013</td>
</tr>
<tr>
<td>path1-2</td>
<td>Rv2610c</td>
<td>0.04</td>
</tr>
<tr>
<td>path1-2</td>
<td>Rv0324</td>
<td>0.044</td>
</tr>
<tr>
<td>path1-3</td>
<td>Rv3133c</td>
<td>2.30E-10</td>
</tr>
<tr>
<td>path1-3</td>
<td>Rv2359</td>
<td>2.47E-03</td>
</tr>
<tr>
<td>path1-3</td>
<td>Rv3066</td>
<td>2.47E-03</td>
</tr>
<tr>
<td>path1-3</td>
<td>Rv0967</td>
<td>0.027</td>
</tr>
<tr>
<td>path2-1</td>
<td>Rv3597c</td>
<td>2.72E-03</td>
</tr>
<tr>
<td>---------</td>
<td>---------</td>
<td>----------</td>
</tr>
<tr>
<td>path2-1</td>
<td>Rv0602c</td>
<td>0.012</td>
</tr>
<tr>
<td>path2-1</td>
<td>Rv0653c</td>
<td>0.017</td>
</tr>
<tr>
<td>path2-2</td>
<td>Rv1049</td>
<td>4.81E-04</td>
</tr>
<tr>
<td>path3-1</td>
<td>Rv3066</td>
<td>2.11E-03</td>
</tr>
<tr>
<td>path3-1</td>
<td>Rv3574</td>
<td>0.01</td>
</tr>
<tr>
<td>path3-1</td>
<td>Rv0081</td>
<td>0.019</td>
</tr>
<tr>
<td>path3-1</td>
<td>Rv0465c</td>
<td>0.023</td>
</tr>
<tr>
<td>path3-1</td>
<td>Rv1956</td>
<td>0.043</td>
</tr>
<tr>
<td>path3-2</td>
<td>Rv3219</td>
<td>3.52E-07</td>
</tr>
<tr>
<td>path3-2</td>
<td>Rv0967</td>
<td>8.01E-04</td>
</tr>
<tr>
<td>path3-2</td>
<td>Rv2359</td>
<td>0.022</td>
</tr>
<tr>
<td>path2-1-1</td>
<td>Rv3133c</td>
<td>4.30E-04</td>
</tr>
<tr>
<td>path2-1-1</td>
<td>Rv0081</td>
<td>5.77E-03</td>
</tr>
<tr>
<td>path2-1-1</td>
<td>Rv0494</td>
<td>6.06E-03</td>
</tr>
<tr>
<td>path2-1-1</td>
<td>Rv3597c</td>
<td>0.03</td>
</tr>
<tr>
<td>path2-1-2</td>
<td>Rv3574</td>
<td>0.03</td>
</tr>
<tr>
<td>path2-2-1</td>
<td>Rv3249c</td>
<td>6.78E-03</td>
</tr>
<tr>
<td>path2-2-2</td>
<td>Rv0494</td>
<td>0.024</td>
</tr>
<tr>
<td>path3-1-1</td>
<td>Rv1255c</td>
<td>3.48E-03</td>
</tr>
<tr>
<td>path3-1-1</td>
<td>Rv0602c</td>
<td>0.011</td>
</tr>
<tr>
<td>path3-1-1</td>
<td>Rv3133c</td>
<td>0.016</td>
</tr>
<tr>
<td>path3-1-1</td>
<td>Rv1985c</td>
<td>0.021</td>
</tr>
<tr>
<td>path3-1-1</td>
<td>Rv1994c</td>
<td>0.021</td>
</tr>
<tr>
<td>path3-1-1</td>
<td>Rv3249c</td>
<td>0.033</td>
</tr>
<tr>
<td>path3-1-1</td>
<td>Rv3246c</td>
<td>0.034</td>
</tr>
<tr>
<td>path3-2-1</td>
<td>Rv1255c</td>
<td>3.74E-03</td>
</tr>
<tr>
<td>path3-2-1</td>
<td>Rv0324</td>
<td>0.011</td>
</tr>
<tr>
<td>path3-2-1</td>
<td>Rv3574</td>
<td>0.021</td>
</tr>
<tr>
<td>path3-2-2</td>
<td>Rv3597c</td>
<td>1.19E-04</td>
</tr>
<tr>
<td>path3-2-2</td>
<td>Rv0494</td>
<td>9.16E-03</td>
</tr>
<tr>
<td>path3-2-2</td>
<td>Rv3246c</td>
<td>0.018</td>
</tr>
<tr>
<td>path3-2-2</td>
<td>Rv0767c</td>
<td>0.026</td>
</tr>
<tr>
<td>path3-2-2</td>
<td>Rv2359</td>
<td>0.048</td>
</tr>
<tr>
<td>path3-1-2-2</td>
<td>Rv0494</td>
<td>3.79E-04</td>
</tr>
<tr>
<td>path3-1-2-2</td>
<td>Rv0081</td>
<td>1.69E-03</td>
</tr>
<tr>
<td>path3-1-2-2</td>
<td>Rv2610c</td>
<td>4.92E-03</td>
</tr>
<tr>
<td>path3-1-2-2</td>
<td>Rv0653c</td>
<td>0.022</td>
</tr>
<tr>
<td>path3-1-2-2</td>
<td>Rv0767c</td>
<td>0.029</td>
</tr>
<tr>
<td>Path Name</td>
<td>TF</td>
<td>P-value</td>
</tr>
<tr>
<td>-----------</td>
<td>----------------</td>
<td>---------</td>
</tr>
<tr>
<td>path1</td>
<td>Rv3133c</td>
<td>0.011</td>
</tr>
<tr>
<td>path1</td>
<td>Rv3574</td>
<td>0.012</td>
</tr>
<tr>
<td>path1</td>
<td>Rv3066</td>
<td>0.016</td>
</tr>
<tr>
<td>path1</td>
<td>Rv2887</td>
<td>0.018</td>
</tr>
<tr>
<td>path2</td>
<td>Rv0757</td>
<td>1.06E-03</td>
</tr>
<tr>
<td>path2</td>
<td>Rv1255c</td>
<td>1.84E-03</td>
</tr>
<tr>
<td>path2</td>
<td>Rv3597c</td>
<td>0.016</td>
</tr>
<tr>
<td>path2</td>
<td>Rv2989</td>
<td>0.019</td>
</tr>
<tr>
<td>path3</td>
<td>Rv3219</td>
<td>2.88E-03</td>
</tr>
<tr>
<td>path3</td>
<td>Rv0494</td>
<td>4.62E-03</td>
</tr>
<tr>
<td>path1-1</td>
<td>Rv0967</td>
<td>9.88E-04</td>
</tr>
<tr>
<td>path1-2</td>
<td>Rv0653c</td>
<td>7.48E-03</td>
</tr>
<tr>
<td>path1-2</td>
<td>Rv1994c</td>
<td>0.032</td>
</tr>
<tr>
<td>path2-1</td>
<td>Rv0324</td>
<td>1.57E-06</td>
</tr>
<tr>
<td>path2-1</td>
<td>Rv0081</td>
<td>1.10E-03</td>
</tr>
<tr>
<td>path2-1</td>
<td>Rv0757</td>
<td>2.56E-03</td>
</tr>
<tr>
<td>path2-1</td>
<td>Rv3597c</td>
<td>4.58E-03</td>
</tr>
<tr>
<td>path2-1</td>
<td>Rv0767c</td>
<td>7.24E-03</td>
</tr>
<tr>
<td>path2-1</td>
<td>Rv2989</td>
<td>8.82E-03</td>
</tr>
<tr>
<td>path2-1</td>
<td>Rv1255c</td>
<td>0.032</td>
</tr>
<tr>
<td>path2-2</td>
<td>Rv2324</td>
<td>0.041</td>
</tr>
<tr>
<td>path3-1</td>
<td>Rv3133c</td>
<td>8.22E-03</td>
</tr>
<tr>
<td>path3-2</td>
<td>Rv2034</td>
<td>5.10E-03</td>
</tr>
<tr>
<td>path3-2</td>
<td>Rv1049</td>
<td>6.62E-03</td>
</tr>
<tr>
<td>path3-2</td>
<td>Rv2324</td>
<td>0.043</td>
</tr>
<tr>
<td>path3-2</td>
<td>Rv0081</td>
<td>0.045</td>
</tr>
<tr>
<td>path3-1-1</td>
<td>Rv0967</td>
<td>0.028</td>
</tr>
<tr>
<td>path3-2-1</td>
<td>Rv0967</td>
<td>1.62E-03</td>
</tr>
<tr>
<td>path3-2-2</td>
<td>Rv0465c</td>
<td>0.039</td>
</tr>
<tr>
<td>path3-2-3</td>
<td>Rv0324</td>
<td>0.019</td>
</tr>
<tr>
<td>path3-2-3</td>
<td>Rv0767c</td>
<td>0.045</td>
</tr>
</tbody>
</table>
Bibliography


Curriculum Vitae

Personal Information
Name: Irina Glotova
Office address: 36 Cummington Street, Boston, MA, 02215
Telephone number: (617) 358-6262
Email: iglotova@bu.edu
Date of birth: 30-04-1985
Nationality: Bulgarian

Education
2009 – present  Boston University, Boston
PhD in Bioinformatics (defended on 8/29/13)
2004 – 2009  Lomonosov Moscow State University, Moscow
B.S. in Bioengineering and Bioinformatics
1999 – 2004  Mathematical School, Plovdiv
High school

Research Experience

James Galagan’s Lab, September 2010 – August 2013
“Reconstruction of lipid metabolism regulatory network in Mycobacterium tuberculosis” (PhD thesis). Build hypoxia and macrophage temporal expression models using the DREM algorithm, integrating time course expression data and ChIP-Seq binding. Identified a Rv0081 as a candidate high level regulator broadly predictive of the overall expression of sets of genes during hypoxia and re-aeration. PhoP is predicted to be the main regulator of the genes involved in the synthesis of SL1 and PAT/DAT as a result of pH change which takes place in the macrophage environment but not during hypoxia.

Simon Kasif’s Lab, October 2009 – January 2010
Identified the previously discovered unexpected pattern in the length bias of tandem repeats of length 3 in the reference human genome as statistically significant in other human and eukaryotic genomes. Analyzed repeats of length 4 and 5 as well.

*Challenge Project of the Bioinformatics BU Program, October 2009 – July 2010*

My colleague and I predicted novel regulatory interactions in mouse liver based on the integration of data from expression microarrays, ChIP-Seq and DNase hypersensitivity sites.

*Mikhail Gelfand’s Lab, October 2008 – May 2009*

“Evolution of the bacterial genomes of Yersinia and Erwinia sp.” (specialist thesis) Did an analysis of orthologous groups of genes to identify differences in the metabolic pathways between different species of Yersinia and Erwinia bacteria.

*Vassily Lyubetsky’s Lab, April 2007 – September 2008*

Applied a model of classical attenuation regulation developed in the lab on a number of genomes of Proteo- and Actinobacteria to predict the existence of such regulation in the leader regions of these genomes. Worked with RNA secondary structures.

*LIAFA Laboratory of Algorithmic Informatics, June 2008 [internship]*

Worked on an algorithm for RNA secondary structure prediction.

*Leiden University Medical Center, July 2007 [internship]*

Developed with a colleague a Java application for visualization of the statistical analysis of multiple microarray datasets generated in the medical center.

*IGLA-3D project, Belozersky Institute of Physico-Chemical Biology, June 2006 – March 2007*

Developed a modular algorithm for pairwise three-dimensional protein structure alignment implemented in C.

*Research Interests*

Comparative genomics, systems biology, regulatory network predictions and modeling, data mining.

*Awards*

Boston University Research Fellowship, September 2009 - August 2010.
Software Skills

Programming Languages
Perl, Matlab, C, Java, Python, MySQL, HTML

Microsoft Office
Word, Excel, PowerPoint

Adobe Suite
Photoshop, Illustrator

Experimental Skills
Bacterial cloning and knock-outs, PCR, microplate reader

Language Skills
Bulgarian (fluent), Russian (fluent), English (fluent), German (basic)

Personal
Graphic and web design, photography, painting/drawing, snowboarding, tennis, badminton

Publications

Unexpected properties of genomic tandem repeats. Irina Glotova, Michael Molla, Arthur Delcher, Simon Kasif (to be submitted to Biology Direct as a discovery note).

Poster Presentations


