UNIVERSITY OF CALIFORNIA SAN DIEGO

Modular Mycobacterium: Decomposition of Mycobacterium Tuberculosis RNA-Seq Data through Independent Component Analysis

A thesis submitted in partial satisfaction of the requirements for the degree Master of Science

in

Bioengineering

by

Reo Sungil Yoo

Committee in charge:

Bernhard ø. Palsson, Chair
Gert Cauwenberghs
Victor Nizet
Anand Sastry

2021

Thesis Approval Page

The Thesis of Reo Sungil Yoo is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

University of California San Diego

2021

Table of Contents

List of Supplemental Files

Supplementary File 1: Dataset_Citations.xlsx

Supplementary File 2: TRN_Citation.xlsx

Supplementary File 3: Carbon_Source_Shift.xlsx

Supplementary File 4: Lipid_Core.xslx

Supplementary File 5: Hypoxia_timecourse.xlsx

Supplementary File 6: Virulence.xlsx

List of Figures

Acknowledgements

I would like to thank Professor Bernhard O. Palsson and Dr. Anand Sastry for giving me the opportunity to work on this project, as well as providing mentorship throughout this process.

I would like to thank the members of the Systems Biology Research Group for their help as well throughout this process, especially Saugat Poudel, Siddharth Chauhan, Tahani Bulushi, Annie Yuan, and Nitasha Menon. I would also like to thank Dr. Daniel Zielinski and Zack Haiman for helping me get started in the lab.

The methods section, in part, has been submitted for publication of the material as it may appear in "*Mining public expression databases to extract microbial transcriptional regulatory networks*", 2021, Anand V. Sastry, Saugat Poudel, Kevin Rychel, Cam Lamoureux, Siddharth Chauhan, Zachary Haiman, Yara Seif, Tahai Al Bulushi, and Bernhard O. Palsson. The thesis author was a contributing investigator to this paper.

Abstract of Thesis

Modular Mycobacterium: Decomposition of Mycobacterium Tuberculosis RNA-Seq Data through Independent Component Analysis

by

Reo Sungil Yoo

University of California San Diego, 2021

Professor Bernhard ø. Palsson, Chair

Mycobacterium Tuberculosis is an infectious disease and a serious public health concern due to the organism's adaptive transcriptional response to environmental stresses via the transcriptional regulatory network (TRN). (Galagan et al., 2013) While many studies seek to better characterize specific portions of the M. tuberculosis TRN, a systems level characterization and analysis of interactions between the controlling transcription factors has yet to be done. Here, we utilize unsupervised machine learning to compartmentalize and describe the transcription factors and regulatory interactions of M. tuberculosis's TRN, allowing us to create a model for how the bacterium responds to environmental stresses.(Boot et al., 2018; Serafini et al., 2019) By applying Independent Component Analysis (ICA) to over 650 transcriptomic samples, we obtained 80 independently regulated gene sets known as "I-modulons" that help to explain the variance in the organisms transcriptional response. This ICA structure helps to elucidate the function of previously undescribed regulons, as well as the transcriptional shifts that occur

during environmental changes such as shifting carbon sources, oxidative stress, and virulence events. Additionally, this analysis has also uncovered an inherent cluster of transcriptional regulons that connects several important metabolic systems, including lipid catalysis, cholesterol catalysis, and sulfur metabolism. This system-wide analysis of the organism's TRN can help inform future research on effective ways to study and manipulate the transcriptional regulation of *M. Tuberculosis.*

The methods section, in part, has been submitted for publication of the material as it may appear in "*Mining public expression databases to extract microbial transcriptional regulatory networks*", 2021, Anand V. Sastry, Saugat Poudel, Kevin Rychel, Cam Lamoureux, Siddharth Chauhan, Zachary Haiman, Yara Seif, Tahai Al Bulushi, and Bernhard O. Palsson. The thesis author was a contributing investigator to this paper.

Chapter 1: Introduction

1.1: Introduction

Mycobacterium tuberculosis is a dangerous pathogen that is the leading cause of death from a single infectious agent and one of the top 10 causes of death worldwide (World Health Organization, 2020). The pathogen is unique for its impermeable cell wall structure and its ability to infiltrate macrophages, which allows the bacillus to remain dormant for months and even years after the initial infection (Delogu et al., 2013). The evolutionary success of *M. tuberculosis* is also due to its adaptability to various environments, and this adaptability is driven in part by the transcriptional regulatory network (TRN) (Ehrt & Schnappinger, 2007; Galagan et al., 2013; Turkarslan et al., 2015). The TRN helps organize the expression of genes across various environmental conditions such as hypoxia, starvation, oxidative stress, and virulence. Given the complexity of *M. tuberculosis'* response to its environments, a strong understanding of the TRN is required. Additionally, given the impact that *M. tuberculosis* has on global health, new methods to better understand the TRN now can help combat the impact of the pathogen in the near future.

One such method that has been used with great success in other organisms is the decomposition of a compendium of RNA-sequencing data (RNA-seq) expression profile utilizing independent component analysis (Poudel et al., 2020; Rychel et al., 2020; Sastry et al., 2019). ICA decomposition of the data has been shown to consistently capture the underlying transcriptional regulator structure utilizing a set of independently modulating genes known as iModulons. This type of ICA analysis has already been performed on *E. coli*, *S. aureus*, and *B. subtilis*, and has revealed important new discoveries in these organisms. These discoveries include simplification of the complex TRN, functional discovery of new transcription factors, and quantification of transcriptional changes associated with shifting environments (Poudel et al., 2020; Rychel et al., 2020; Sastry et al., 2019). Previous studies have already established ICA as a robust method of capturing the structure of the transcriptional regulatory network (TRN) using independently modulated gene clusters, called iModulons (Poudel et al., 2020; Rychel et al., 2020; Sastry et al., 2019). However, unlike previously defined regulons, which utilize molecular measurement techniques to establish transcriptional groups, iModulons are driven purely by statistical decomposition of

the data. While this statistical approach is useful for providing structure to the complex interactions between transcription factors (TFs), the ICA algorithm can be limited by the volume of data found in the original dataset (Sastry et al., 2021).

In order to gain deeper insight into the structure and operation of the M. tuberculosis TRN, we performed the same ICA decomposition. We compiled 657 high quality RNA-seq expression profiles from a variety of publicly available datasets on the NCBI Sequence Read Archives for this analysis, and extracted 80 robust iModulons. We then utilized iModulons to accelerate discovery in M. tuberculosis by: 1) quantitatively describing the organization of the TRN, 2) elucidating the function of new transcription factors, 3) defining transcriptional shifts that occur across changes in carbon sources, oxygen levels, and virulence states, 4) clustering various transcription factors into a core stress response (Kodama et al., 2012).

Chapter 2: Methods

The functions used in this study and description of the methods for pulling and processing RNA-Seq data, running ICA, and computing iModulon enrichments were adapted from the Pymodulon methods paper from Sastry et al. (unpublished).

2.1. Compiling all public transcriptomics data for an organism

The NCBI Sequence Read Archive (SRA) is a public repository for sequencing data that is partnered with the EMBL European Nucleotide Archive (ENA), and the DNA Databank of Japan (DDBJ) (Kodama et al., 2012). We provide a script (https://github.com/avsastry/nf-rnaseq-bacteria/tree/main/download_metadata) that uses Entrez Direct (Kans, 2020) to search for all public RNA-seq datasets and compile the metadata into a single tab-separated file. Each row in the file corresponds to a single experiment, and users may manually add private datasets.

2.2. Processing prokaryotic RNA-seq data

The tab-separated metadata file can be directly piped into the prokaryotic RNA-seq processing pipeline. This pipeline is implemented using Nextflow v20.01.0 (Di Tommaso et al., 2017) for reproducibility and scalability, and is available at https://github.com/avsastry/nf-rnaseq-bacteria. To process the complete *M. tuberculosis* RNA-seq dataset, we used Amazon Web Services (AWS) Batch to run the Nextflow pipeline.

The first step in the pipeline is to download the raw FASTQ files from NCBI using fasterq-dump (https://github.com/ncbi/sra-tools/wiki/HowTo:-fasterq-dump). Next, read trimming is performed using Trim Galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) with the default options, followed by FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) on the trimmed reads. Next, reads are aligned to the genome using Bowtie (Langmead et al., 2009). The read direction is inferred using RSEQC (Wang et al., 2012) before generating read counts using featureCounts (Liao et al., 2014). Finally, all quality control metrics are compiled using MultiQC (Ewels et al., 2016) and the final expression dataset is reported in units of log-transformed Transcripts per Million (log-TPM).

2.3. Quality Control and Data Normalization

To guarantee a high quality expression dataset for *M. tuberculosis*, data that failed any of the following four FASTQC metrics were discarded: per base sequence quality, per sequence quality scores,

per base n content, and adapter content. Samples that contained under 500,000 reads mapped to coding sequences were also discarded. Hierarchical clustering was used to identify samples that did not conform to a typical expression profile, as these samples often use non-standard library preparation methods, such as ribosome sequencing and 3' or 5' end sequencing (Ziemann et al., 2019).

Manual metadata curation was performed on the data that passed the first four quality control steps. Information including the strain description, base media, carbon source, treatments, and temperature were pulled from the literature. Each project was assigned a short unique name, and each condition within a project was also assigned a unique name to identify biological and technical replicates. After curation, samples were discarded if (a) metadata was not available, (b) samples did not have replicates, or (c) the Pearson R correlation between replicates was below 0.95. Finally, the log-TPM data within each project was centered to a project-specific reference condition.

2.4. Computing the optimal number of robust Independent Components

To compute the optimal independent components, an extension of ICA was performed on the RNA-seq dataset as described in McConn et al. (unpublished).

Briefly, the scikit-learn (v0.23.2) (Pedregosa et al., 2011) implementation of FastICA (Hyvärinen, 1999) was executed 100 times with random seeds and a convergence tolerance of 10^{-7}. The resulting independent components (ICs) were clustered using DBSCAN (Ester et al., 1996) to identify robust ICs, using an epsilon of 0.1 and minimum cluster seed size of 50. To account for identical with opposite signs, the following distance metric was used for computing the distance matrix:

$$d_{x,y} = 1 - ||\rho_{x,y}||$$

where $\rho_{x,y}$ is the Pearson correlation between components *x* and *y*. The final robust ICs were defined as the centroids of the cluster.

Since the number of dimensions selected in ICA can alter the results, we applied the above procedure to the *M. tuberculosis* dataset multiple times, ranging the number of dimensions from 10 to 260 (i.e. the approximate size of the dataset) with a step size of 10. To identify the optimal dimensionality, we compared the number of ICs with single genes to the number of ICs that were correlated (Pearson R > 0.7) with the ICs in the largest dimension (called "final components"). We selected the number of

dimensions where the number of non-single gene ICs was equal to the number of final components in that dimension.

2.5. Compiling gene annotations

The gene annotation pipeline can be found at https://github.com/SBRG/pymodulon/blob/master/docs/tutorials/creating_the_gene_table.ipynb. Gene annotations were pulled from AL009126.3. Additionally, KEGG (Kanehisa et al., 2021) and Cluster of Orthologous Groups (COG) information were obtained using EggNOG mapper (Huerta-Cepas et al., 2017). Uniprot IDs were obtained using the Uniprot ID mapper (UniProt Consortium, 2021), and operon information was obtained from Biocyc (Karp et al., 2019). Gene ontology (GO) annotations were obtained from AmiGO2 (The Gene Ontology Consortium, 2019). The known transcriptional regulatory network was obtained primarily from the Galagan and MTB Network portal databases. (Galagan et al., 2013; Turkarslan et al., 2015)

2.6. Computing iModulon enrichments

iModulon enrichments against known regulons were computed using Fisher's Exact Test, with the false discovery rate (FDR) controlled at 10^{-5} using the Benjamini-Hochberg correction. Fisher's Exact Test was used to identify GO and KEGG annotations as well, with an FDR < 0.01.

Additional functions for gene set enrichment analysis are located in the *enrichment* package, including a generalized gene set enrichment function and an implementation of the Bonferroni-Hochberg false discovery rate (FDR).

2.7. Calculating Differentially Expressed iModulons Across Conditions

The difference in activity of iModulons were compared across relevant conditions and significantly changed iModulons were calculated utilizing a lognormal probability distribution. The probability of obtaining the observed difference was calculated and the FDR was calculated. iModulon changes were considered significant if the difference was greater than 5 and FDR < .01.

DIMA scatter plots plot the activities of iModulons under one condition versus another, and allow for the visualization of significantly changed iModulons. 1D DIMA plots plot iModulons under one condition to a reference condition. Reference conditions have been normalized to have 0 activity across all iModulons, and thus a bar plot is used instead of a scatter plot.

2.8. Biological Analysis

Analysis of iModulon activities and determination of physiological meaning was performed utilizing several methods. The algorithms used to obtain the results found in this paper as well as create the accompanying figures can be found at the project's Github page (https://github.com/Reosu/modulome_mtb/tree/master/notebooks)

2.9. Calculating Clusters of iModulons by Activity

The activities of iModulons were first vectorized and then clustered using Seaborn clustermap. (Waskom et al., 2015) Pearson R correlation was used as a distance metric, and pairwise distances for each iModulon were calculated. After creation of the clustermap, the scikit-learn agglomerative clustering function was performed on the clustermap. (Pedregosa et al., 2011) Optimal cluster sizes were obtained by computing the varying the threshold statistic for agglomerative clustering and finding the optimal silhouette score. Once iModulons clusters were calculated, clusters that had above average Pearson R statistics were manually inspected to determine physiological function.

2.10 Acknowledgments

The methods section, in part, has been submitted for publication of the material as it may appear in "*Mining public expression databases to extract microbial transcriptional regulatory networks*", 2021, Anand V. Sastry, Saugat Poudel, Kevin Rychel, Cam Lamoureux, Siddharth Chauhan, Zachary Haiman, Yara Seif, Tahai Al Bulushi, and Bernhard O. Palsson. The thesis author was a contributing investigator to this paper.

Chapter 3: Results

3.1. Independent component analysis of public access data reveals the structure of the M. tuberculosis transcriptional regulatory network

In order to capture the complete spectrum of *M. tuberculosis* transcriptional response, we took advantage of the publicly available data found in NCBI's Sequence Read Archive (SRA) and obtained 980 RNA-seq expression profiles from 53 separate studies(Kodama et al., 2012). Each sample was processed through a standardized pipeline (citation pending) to assess the dataset quality and filter out poor quality datasets (See methods). The final dataset was composed of 657 samples, spanning various conditions that capture *M. tuberculosis*'s response to various nutrient sources, stressors, antibiotics, and virulence events. After the final dataset was obtained, a previously developed ICA algorithm was used to decompose the data into 80 robust iModulons. (Sastry et al., 2019)

In order to provide biological meaning to the new clusters, robust iModulons were categorized by mapping each gene cluster to known TFs, KEGG pathways, GO terms, or other knowledge clusters proposed by literature. An iModulon was considered mapped to a particular knowledge cluster if there was a statistically significant (FDR < .01) overlap between the genes found in the iModulon and the genes found within the classifier. Some iModulons were manually annotated due to shared functions of constituent genes, or presence of knocked-out genes. ICA also captured the activity of each iModulon across experimental conditions, and these were used to examine the response of *M. tuberculosis* to various environments. In order to minimize batch effects, activity levels for each project were centered to a reference condition within the experimental subset (Sastry et al., 2021). After examining the mapped classifier and activities, each iModulon was assigned a functional category (Figure 1.1). While most categories indicated biological function, some categories indicated specific properties of the dataset. For example, the Unknown Function category contains iModulons that have been mapped to an established TF regulon, but the function of the TF remains unclear. Uncharacterized iModulons are those which had little overlap with known TFs or classifiers, but still contained a significant number of genes. Finally, Single Gene iModulons are those which track the expression of a single gene, and are treated as an artifact of the ICA decomposition (citation pending).

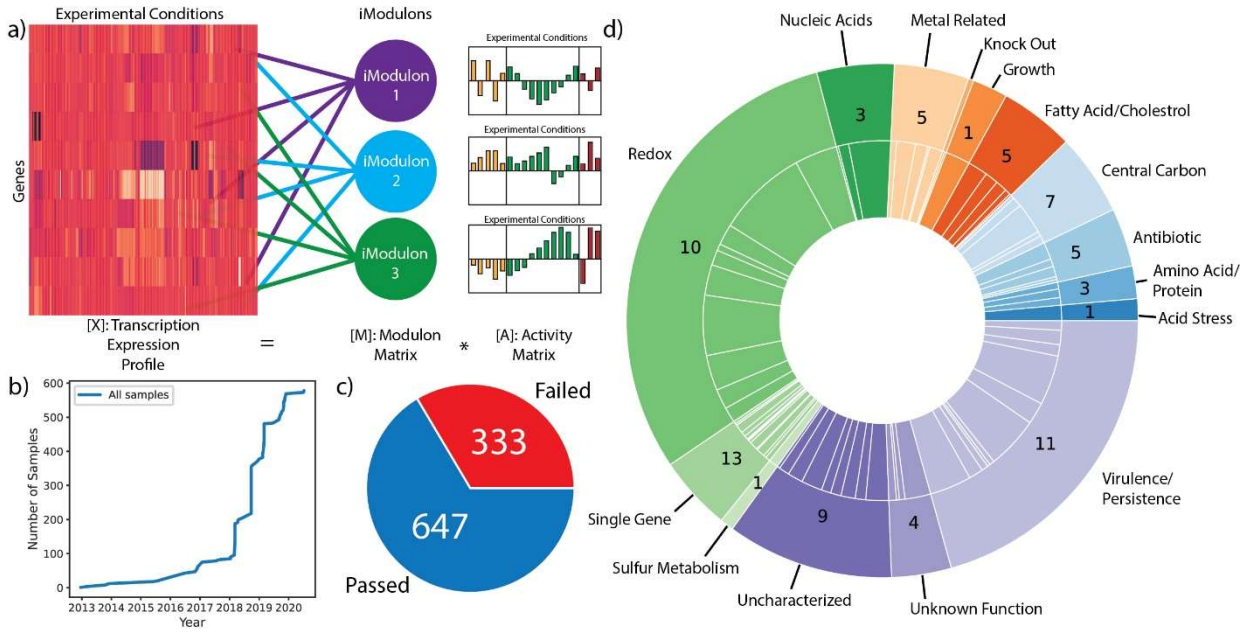

Figure 1.1: QC/QA, ICA Decomposition, and iModulon Characterization of M. Tuberculosis RNA-seq Data from Sequence Read Archive (a) iModulons are clusters of genes computed by decomposing RNA-Seq data into independently modulated sets. ICA decomposition of an RNA-Seq expression (X) matrix will create the gene weight (M) matrix and the activity (A) matrix. The M matrix defines the genes within an iModulon, and the activity matrix corresponds to the relative expression of the iModulon across the experimental conditions (e.g different media, hypoxia, growth in macrophages, etc.) (b) A timeline of the number of samples used in this study added to the Sequence Read Archive. (c) Percentage of samples with metadata that passed and failed the QC/QA process. The RNA-seq data and associated metadata from 980 H37Rv SRA samples were processed, and 647 samples passed all QC/QA metrics. (d) A donut chart of the 80 M. tuberculosis iModulons. The inner ring indicates the size and function of each iModulon. The outer ring represents the functional categories for the iModulons, and the number represents how many iModulons belong to that category.

3.2. Modulons Capture Activity of Known Transcriptional Regulators Zur and Lsr2

To demonstrate how iModulons capture the structure and biology of the *M. tuberculosis* TRN, two iModulons associated with well-characterized TFs were examined: Zur and Lsr2.

Zur, previously known as FurB, is a zinc uptake regulator that modulates the expression of ribosomal proteins, zinc transporters, and the metal homeostasis secretory system ESX-3 (Maciąg et al., 2007; Pandey et al., 2015). It is known that Zur activity is influenced by molecular zinc concentration and has been associated with virulence (Maciąg et al., 2007). The relation between zinc transport and virulence is likely due to the ability of host cells to sequester metal ions such as iron and zinc to deprive parasites of essential nutrients (Cellier, 2012). As such, pathogenic bacteria have evolved efficient secretory systems to efficiently capture required metals (Neyrolles et al., 2015). In *M. tuberculosis*, ESX-3 is an example of such a secretory system, and is responsible for secreting factors such as the PE and PPE family of proteins and iron chelators (Tufariello et al., 2016). We found that the Zur iModulon captured many of the genes found in the Zur regulon, including ESX-3 (pie chart?). Additionally, the iModulon was highly upregulated in macrophage infection conditions when compared to non-virulent controls, validating that the Zur iModulon reflects the activity of the Zur TF. (Fig XX) (Peterson et al., 2019). Interestingly, while Zur is usually activated by zinc ions, the Zur iModulon exhibited high activities in both high and low iron concentrations. This would suggest that Zur may is also be sensitive to iron ions and may help to establish iron homeostasis together with the iron uptake regulator, IdeR. (Rodriguez et al., 2002)

The Lsr2 TF is a small, basic protein that has been found to bind to DNA in a sequence-independent manner (Gordon et al., 2008). Lsr2 acts as a global transcriptional repressor that controls the expression of genes during virulence, adaptations to oxygen levels, and DNA organization (Bartek et al., 2014; Kołodziej et al., 2021). To validate that the Lsr2 iModulon reflects the known biology, we examined the iModulon activity under virulence and hypoxia conditions. Under virulence conditions, it was found that Lsr2 may have two distinct responses based on the type of host cell. In conditions of *in vivo* infections of mice neutrophils, the activity of the Lsr2 iModulon significantly decreased (Figure 2.1). However, during infections of mice macrophages (bone marrow derived or THP-1), the iModulon had significantly increased activity (Mishra et al., 2019; Peterson et al., 2019). This confirmed that the

iModulon reflected the virulence ability of the TF, but also suggested that Lsr2 regulation is dependent on the host cell type. We also confirmed that the iModulon was activated during hypoxia conditions, which reflects the expected behaviour of the TF (Peterson et al., 2019).

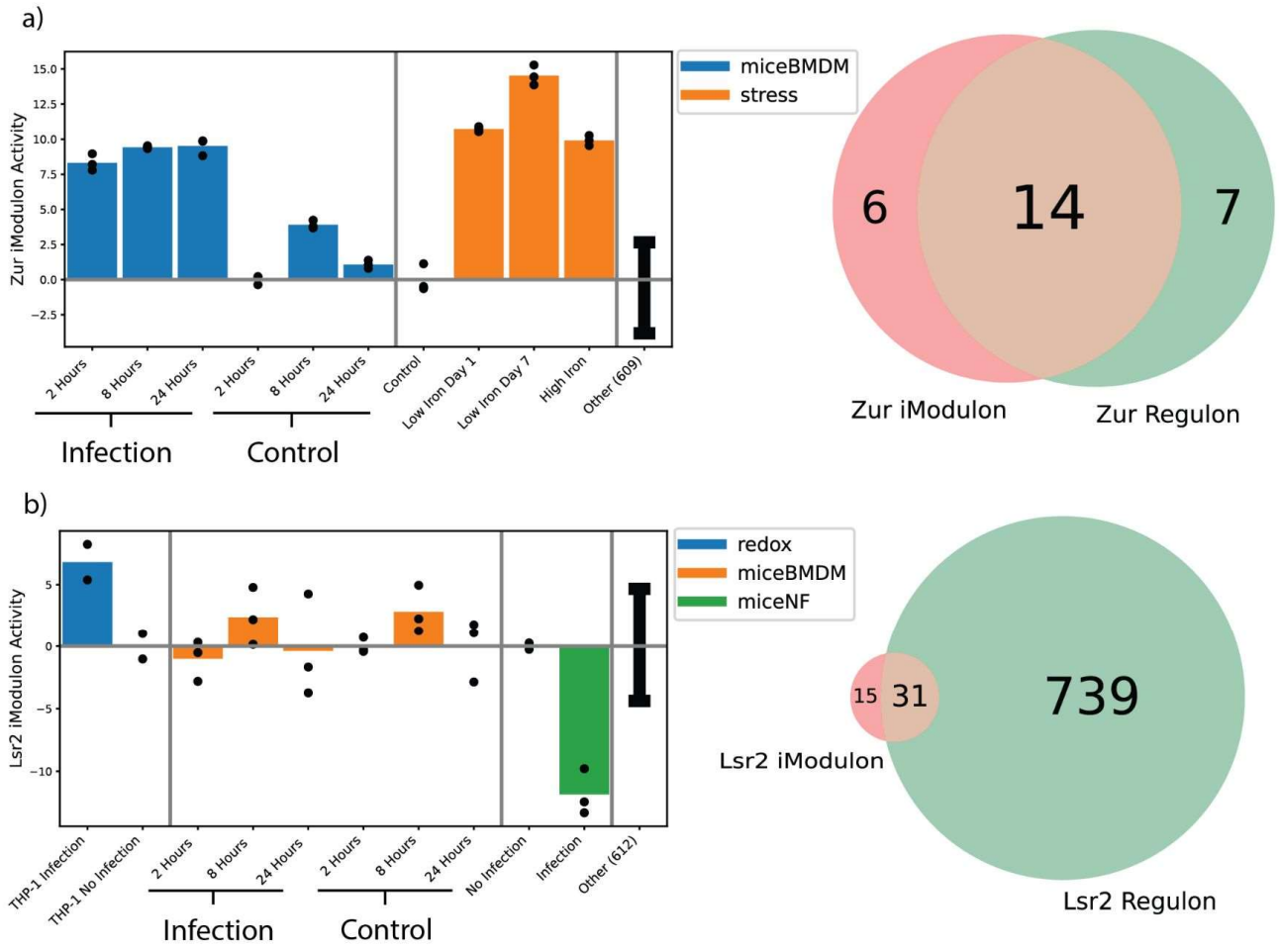

Figure 2.1: iModulons Capture Activity of Known Transcriptional Regulators Zur and Lsr2
For activity bar plots, error bars represent mean and standard deviation of all other samples, black dots represent the activity of each replicate for a condition, and vertical gray bars separate the samples into projects. Each project is normalized to a reference condition within that project such that the reference condition represents zero activity. (a) Left: Bar plot representing the activity of the Zur iModulon across virulence and iron conditions; Right: Venn diagram showing the genes that overlap between the established Zur regulon and the calculated iModulon (b) Left: Bar plot representing the activity of the Lsr2 iModulon across three different virulence conditions (THP-1 macrophages, bone marrow derived macrophages, and neutrophils); Right: Venn diagram showing the genes that overlap between the established Lsr2 regulon and the calculated iModulon

3.3. iModulons Reveal New Functions in both Characterized and Novel Transcription Factors

Since iModulons captured the structure and function of *M. tuberculosis* transcriptional regulation, we investigated if iModulons could be used for the discovery of new TF functions. Here, we selected two iModulons, PhoP and Rv0681 and examined their activities.

PhoP is part of the two-component system PhoPR, which controls genes responsible for hypoxia adaptation, lipid metabolism, respiration, and various stress responses (Pérez et al., 2001). The PhoP regulon also controls the expression of the redox responsive TF WhiB3, which is essential for virulence (Bansal et al., 2017; Gonzalo-Asensio et al., 2008; Pérez et al., 2001). Within our iModulon structure, we found one iModulon with statistically significant overlap to the PhoP regulon, and the activity was examined during acidic exposure. Under acidic environments (pH 4.5), the PhoP iModulon was significantly upregulated when compared to bacteria exposed to a neutral environment (pH 6.6), which agrees with previous experiments (Feng et al., 2018). However, the same acidic exposure study in combination with iModulons revealed a new possible regulatory interaction for PhoP. When a *whiB3* knockout strain was also exposed to the same acidic and neutral pH levels, the PhoP iModulon was significantly downregulated compared to the wild type. (Figure 3.1 a) This would suggest that while PhoP controls the expression of *whiB3*, WhiB3 may simultaneously control the expression of *phoP,* which has not been reported before.

To determine that the decrease in activity was a feature of the biological data and not an artifact from the knockout of one gene, the normalized gene expression levels of genes within the PhoP iModulon were examined. Five genes were selected, including both *phoP* and *whiB3*, and their expression levels were examined. We found that all genes except phoP had significantly decreased expression levels in the *whiB3* knockout strain in both acidic and neutral condition. (Figure 3.1 b) We also examined the WhiB3 regulon to determine if WhiB3 itself regulates any of the genes in the PhoP iModulon, which would explain why the knockout of *whiB3* would result in decreased activity in the iModulon. However, only one gene was found to overlap with both the previously established WhiB3 regulon and the PhoP iModulon, which does not fully explain the decreased expression of genes in the *whiB3* mutant strain. (Figure 3.1 c) Taken together, we hypothesize that WhiB3 does indeed have a significant regulatory effect on *phoP*,

though the mechanisms remain unknown. Interestingly, the expression level of *phoP* was relatively unchanged in the whiB3 knockout strain.

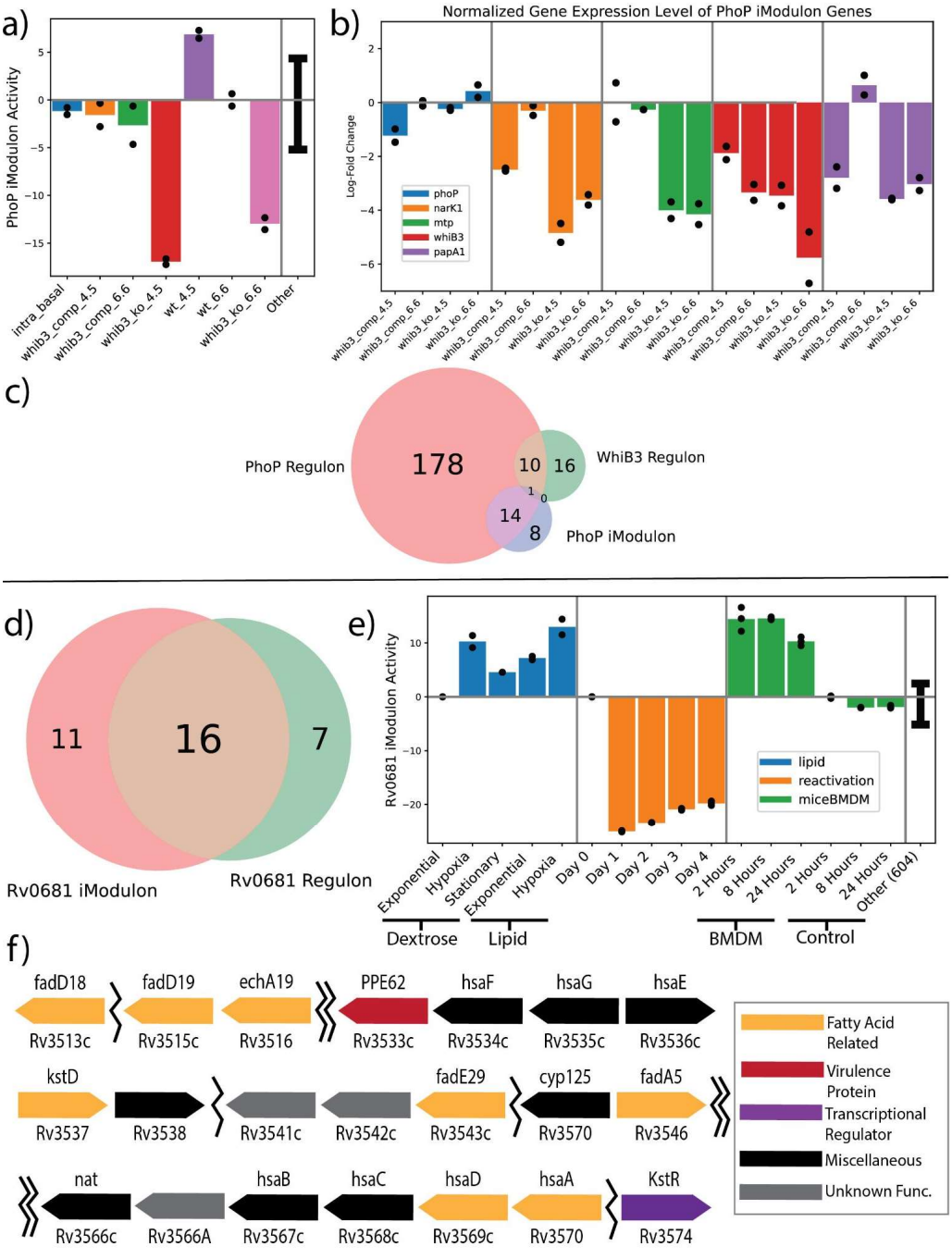

Figure 3.1: Discovery of new TRN interactions for PhoP and Rv0681 (a) Bar plot representing the activity of the PhoP iModulon across redox and WhiB3 knockout conditions. (b) Bar plot representing the normalized gene expression of several genes found within the PhoP iModulon. Expression levels indicate that the relationship between *whib3* knockout and PhoP is a biological feature (c) Venn diagram displaying the genes that overlap between the *phoP* regulon, the *whiB3* regulon, and the PhoP iModulon (d) Venn diagram displaying the genes that overlap between the predicted Rv0681 regulon and the calculated Rv0681 iModulon (e) Barplot displaying the activities of the Rv0618 iModulon across lipid, hypoxic reactivation, and virulence conditions. (f) A diagram that characterizes the position and function of the genes found in the Rv0681 iModulons. Many of these genes are related to fatty acids and cholesterol, including the KstR transcription factor (Kendall et al., 2010)

iModulons can also provide insight into TFs whose functions are unknown. Rv0681 is an uncharacterized HTH-type transcriptional regulator which has been experimentally shown to be phosphorylated by the PknH kinase (Kelkar et al., 2011; Zheng et al., 2007). The Rv0681 iModulon was found to have significant overlap with the predicted Rv0681 regulon, and thus was an ideal candidate for functional discovery (Figure 3.1d). Of the 23 genes predicted or experimentally determined to be bound by Rv0681, 16 genes were found within the iModulon (Galagan et al., 2013; Turkarslan et al., 2015). A large proportion of the genes in the Rv0681 iModulon were labeled with the “lipid transport and metabolism” COG annotation. Additionally, the KstR TF, which is an important regulator for cholesterol metabolism in *M. tuberculosis*, was also found within the iModulon (Kendall et al., 2010). Given the genes found within the iModulon, we hypothesized that Rv0681 may be involved in the regulation of lipid and cholesterol metabolism.

To add additional insight to the possible function of the iModulon, we examined the activity of the iModulon across three projects. In the first project, *M. tuberculosis* was grown on either dextrose or lipid-only media, during exponential phase, stationary phase, or hypoxic exposure(Aguilar-Ayala et al., 2017). We found that when the bacterium was in either the exponential or stationary phase, a switch from a dextrose media to a lipid-only media led to a significant up-regulation in the Rv0681 iModulon activity (Figure 3.1 e).

In the second project, *M. tuberculosis* was first induced into a persistence state via hypoxia. The bacteria was then reactivated via re-aeration, and RNA-Seq was performed once a day for 4 days (Du et al., 2016). The Rv0681 iModulon to had significantly decreased activity when reactivating from dormancy (Figure 3.1 e), suggesting that Rv0681 is important for hypoxia and dormancy response, but is unnecessary when ample oxygen is available.

Due to the close relationship between lipids, hypoxia, and virulence, we examined a third project that tested the infection of mouse bone marrow-derived macrophages (BMDM) (Ehrt & Schnappinger, 2007). The iModulon was significantly upregulated during infections of the macrophage when compared to non-infection controls at all time points, confirming that the iModulon is involved with virulence as well.

Altogether, we propose that Rv0681 is a transcriptional factor that controls lipid metabolism to promote survival in stressful conditions such as hypoxia and infection.

3.4. iModulons Elucidate Transcriptional Responses to Shifts in Carbon Sources

While individual iModulons can provide information about a single TF, one of their most useful functions is to simplify organism-wide transcriptional responses. Thus, we examined the shifts in activities of our iModulons under various environmental conditions to elucidate the system-wide shifts in the TRN. In order to establish a baseline for this type of analysis, we started by examining how the bacterium changes its transcriptional responses under various carbon sources.

In order to study how different carbon sources can affect the TRN, we utilized data obtained from a study where *M. tuberculosis* was fed glucose, lactate, or pyruvate as a sole carbon source (Serafini et al., 2019). In total, the study contained 6 different conditions, representing the 3 carbon source conditions (glucose, lactate, and pyruvate) with two time points each (6 hours and 24 hours). The original study found that the glyoxylate shunt, which plays a role in beta oxidation of lipids, is required for growth in both lactate and pyruvate only media. Additionally, it was also found genes related to the Krebs cycle, such as *pckA* and *icl1*, were highly induced in both lactate and pyruvate conditions, though there were also some genes that were found to be essential in only one of the carbon sources. To assess if iModulons reflect these previous findings, we created several DIMA (Differential iModulon Activity) plots to examine which iModulons were significantly changed between the glucose and the alternate carbon source. Four iModulons were of particular interest: Fumarate Reductase, Sulfur Metabolism, PrpR, and BlaI.

For cells growing on both lactate and pyruvate, the Fumarate Reductase iModulon was up-regulated at all time points compared to the glucose-fed conditions. The Fumarate Reductase iModulon contains genes associated with the TCA cycle and fatty acid synthesis, including *icl2, pckA, and fad* genes (Figure 4.1 b). Many of the genes in this iModulon were also highlighted by the original study for survival in lactate and pyruvate media, which include those genes that regulate the glyoxylate shunt. However, the Fumarate Reductase iModulon also captures the expression dynamics of many other genes not found in the original paper. These include the *fad* genes, which code for various enzymes in fatty acid synthesis, *yrbE1* putative permeases, and the *mce1R* transcription factor, which is a vital regulator for virulence (Casali et al., 2006; Forrellad et al., 2014). Many of these genes are important for maintaining lipid homeostasis, and suggests that the systems that help metabolize pyruvate and lactate are connected to the same systems that metabolize or synthesize lipids. Additionally, the inclusion of the

virulence regulator Mce1R suggests that high levels of lactate and pyruvate may trigger the virulence response in *M. tuberculosis*.

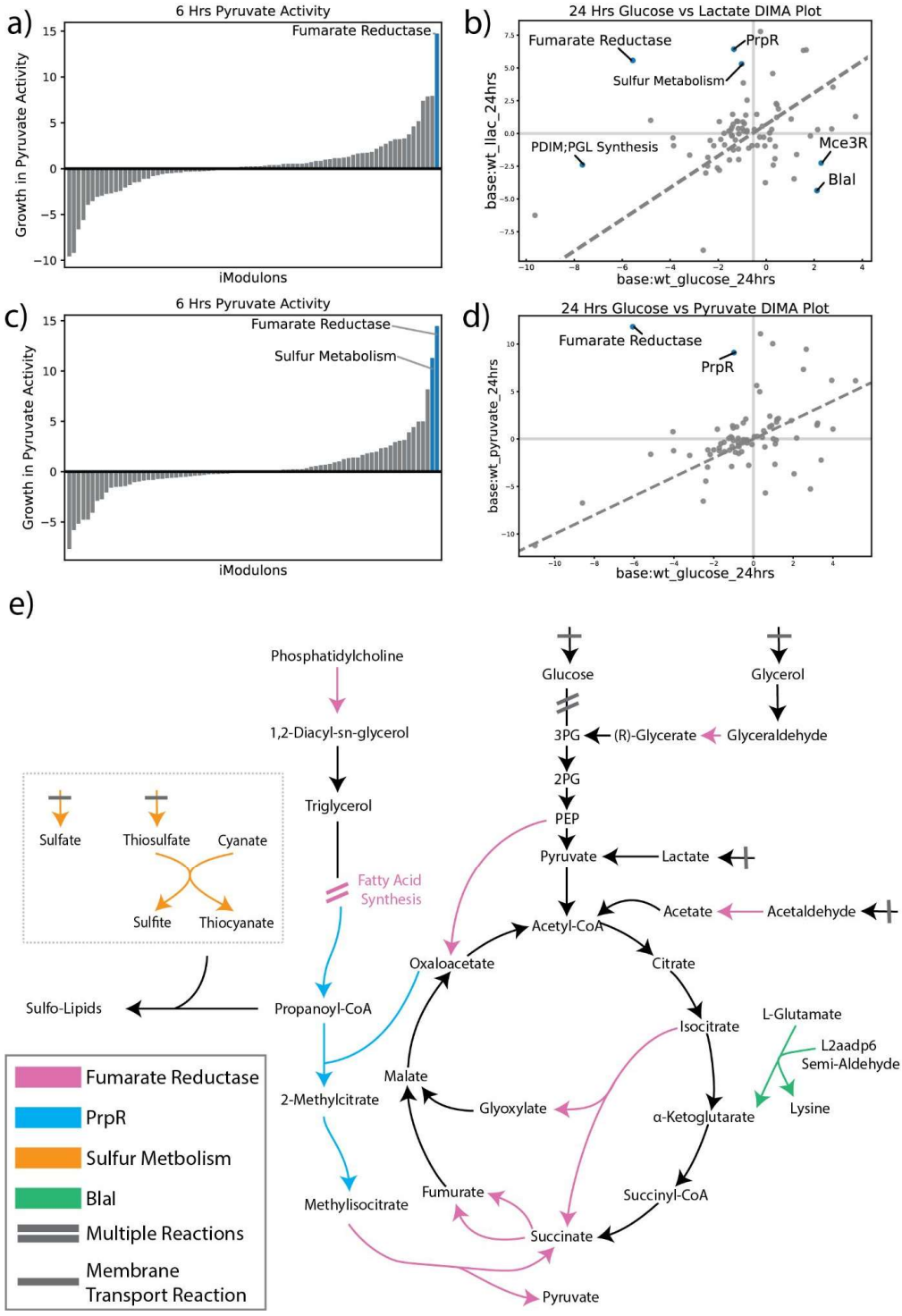

Figure 4.1: iModulons Illuminate Metabolic Shifts from Changes in Carbon Source
(a) DIMA plots representing the differentially expressed iModulons at 6 hours and 24 hours, using either L-lactate or pyruvate as a carbon source. Fumarate Reductase, Sulfur Metabolism, and PrpR are the most consistently upregulated iModulons (b) A metabolic map representing the reactions controlled by differentially expressed iModulons across carbon source shifts. Arrows represent reactions between metabolites, and reactions with bars represent transport from the environment. Map displays how reactions controlled by the significant iModulons are connected to one another.

Further evidence of a connection between alternate carbon sources, fatty acid synthesis, and virulence can be found by the significant upregulation of the Sulfur Metabolism iModulon during growth in L-lactate media. Using the metabolic map, we found that the Sulfur Metabolism iModulon controlled reactions for sulfate and thiosulfate import, as well as the conversion of the transfer of the thiol group from thiosulfate to cyanate (Kavvas et al., 2018). These pathways suggest that the Sulfur Metabolism iModulon is strongly related to the synthesis of sulfolipids. Surface-exposed sulfolipids are a family of sulfated acyl trehalose in mycobacteria that have shown to have virulent properties. While two families of sulfolipids have been identified, labeled as SL-1 and SL-2, the synthesis pathways and virulent properties of the SL-1 family are better characterized (Ly & Liu, 2020). SL-1 sulfolipid synthesis is initiated with the import of sulfate, and through a series of reactions sulfate is converted into trehalose-2-sulfate (TS2). Fatty acids such as palmitoyl-CoA are incorporated into this pathway in order to esterify TS2. After a further acylation of the resulting product, SL-1 sulfolipids are synthesized (Kumar et al., 2007). We theorize that, given the co-expression of the Fumarate Reductase and Sulfur Metabolism iModulons, lactate is a signal to the cell that it is residing in a host cell environment. This induces the cell to not only upregulate pathways to metabolize lactate, but also upregulates the uptake of sulfate and thiosulfates for the synthesis of virulent sulfolipids.

We also found evidence of time-dependent iModulon responses during exposure to alternative carbon sources. At 24 hours, we found significant upregulation of the PrpR iModulon under both lactate and pyruvate conditions. In *M. tuberculosis*, the PrpR TF is responsible for control of the *prp* operon, which codes for several key enzymes that integrate propionyl-CoA into the methylcitrate cycle (Tang et al., 2019). Propionyl-CoA is important to the virulence response (Wilburn et al., 2018), and here it plays a different metabolic role. Many of the genes found in both the PrpR iModulon and TF break down Propionyl-CoA into pyruvate and succinate, which then can be used in the methylcitrate cycle to produce NADH. The appearance of the PrpR iModulon at 24 hours and not at 6 hours suggests that this is a starvation response, and we theorize that the iModulon is activated to supplement the production of NADH and ATP from solely lactate or carbon sources. In addition, the BlaI iModulon was significantly downregulated at 24 hours, but only under lactate conditions. The BlaI TF regulates genes involved in

antibiotic resistance and ATP synthesis (Sala et al., 2009). However, metabolic models of *M. tuberculosis* reveal that the only relevant reaction regulated by BlaI was a conversion of L-glutamate into lysine and alpha-ketoglutarate. While the full purpose of this iModulon in response to alternative carbon sources remains unclear, the presence of this iModulon suggests amino acids may be involved in the synthesis of intermediates.

3.5. Redefining the Core Lipid Response in *M. tuberculosis*

Given the previously discussed role of lipids in the consumption of alternate carbon sources and virulence, we were interested in determining if our iModulons were activated during lipid metabolism. Within our dataset, a study examined the differentially expressed genes between dextrose- and lipid-fed *M. tuberculosis* across 3 metabolic states (exponential growth, stationary phase, hypoxia). The study then defined a “core lipid response”, which contained genes that were found to be differentially expressed between dextrose and lipid media across all three metabolic states. This core lipid response was composed of 6 genes: Rv3161c, Rv3160c, Rv0678, Rv1217c, PPE53 and che1 (Aguilar-Ayala et al., 2017). Given that defining a core lipid response can be crucial for identifying potential targets to combat *M. tuberculosis* infections, we were interested in seeing if the structure of iModulons could be used to define a more comprehensive core lipid response utilizing the same RNA-Seq data.

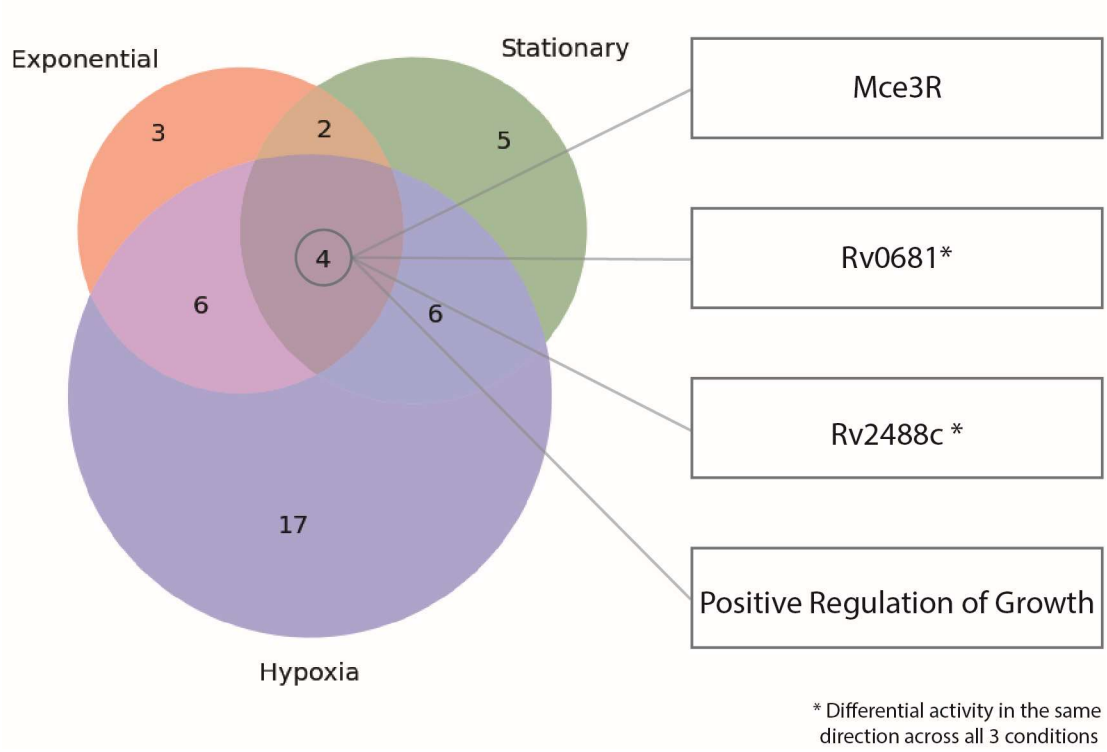

Figure 5.1: The Core Lipid Response can be Expanded using iModulons
A three-way venn diagram was used to identify the core lipid response. 4 iModulons with differential activities across all three cellular states were considered to be part of the core lipid response. Rv0681 and Rv2488c were found to be consistently upregulated across all three cell states, while Mce3R and Positive Regulation of Growth were activated in both directions. Despite the difference in activity direction, we maintain that each of these iModulons are important for survival in lipid rich environments.

In order to determine a more complete core lipid response, the iModulon activities were examined between the same lipid and dextrose conditions. iModulons with significant differential expression across all three metabolic states were labeled as part of the new core lipid response. (Figure 5.1) While the original study identified a core lipid response composed of only 6 genes, our analysis of the same data identified 4 iModulons to be contained within the core lipid response, spanning a total of 80 unique genes. The four iModulons that were found were Mce3R, Rv0681, Rv2488c, and Positive Regulation of Growth (PROG). Surprisingly, we found that while Rv0681 and Rv2488c had consistent upregulation across all three cell states, Mce3R and PROG were found to be both up and downregulated. However, while the reason for this is unclear, we still maintain that all four of these iModulons are important systems for *M. tuberculosis* in a lipid rich environment.

Upon close examination, we found that 5 of the 6 genes previously identified as part of the core lipid response were captured by the Rv2488c iModulon, whereas *che1* was not found in any of the computed iModulons. While we have previously identified Rv0681 as having a lipid related function, the previously uncharacterized Rv2488c iModulon captured most of the genes associated with the original core lipid response. The Rv2488c iModulon primarily regulates the expression of membrane-associated proteins and other transcriptional regulators. One of the most prominent families of proteins within this iModulon were the MmpS/L efflux pumps, which are associated with transporting various metabolites. Interestingly, the specific MmpS/L efflux pumps that were found within the iModulon are implicated with the export of siderophores (Briffotaux et al., 2017). Given the presence of the MmpS/L and other efflux proteins, we theorize that the Rv2488c controls an essential, lipid-activated transcriptional response involved in cellular defense. All together, iModulons provide a clear definition of a core lipid response that adds to our knowledge on how *M. tuberculosis* metabolizes lipids.

We also examined the genes found within the Mce3R and PROG iModulons to determine how they contribute to the core lipid response. We found that the PROG iModulon contains a large set of VapB toxin-antitoxin transcriptional regulators. The VapBC family of proteins are a set of regulators that control cell growth, and there is evidence they play a role in *M. tuberculosis* persistence (Ahidjo et al., 2011). Specifically, the VapB proteins serve as the cognate antitoxin for the VapC toxin, which restricts the growth inhibiting effects of VapC. Though some studies have identified connections between specific

VapBC regulators and lipid metabolism, the implication that the VapBC systems are essential to the metabolism of lipids is novel (Eroshenko et al., 2020). While the Mce3 iModulon, which almost exclusively contains mammalian cell entry proteins, is associated with improving survival in a host cell, it is thought to also serve a role in lipid import (Wilburn et al., 2018). Though this has yet to be confirmed, the presence of this system within the core lipid response adds further evidence to the theory. All together, iModulons provide a clear definition of a core lipid response that adds to our knowledge on how *M. tuberculosis* metabolizes lipids.

3.6. Time-Course Data and iModulons Validate Prior Models of TF Responses to Hypoxia

Hypoxia is an important signal for the TRN of *M. tuberculosis*, as the bacterium enters an altered metabolic state when exposed to reduced oxygen levels *in vitro* and *in vivo* (Galagan et al., 2013; Rustad et al., 2009). While many studies have examined hypoxic response in *M. tuberculosis* utilizing tools such as differential gene expression, we examined whether iModulons could provide additional insight on the organism's hypoxic response.

Here, we analyze the important iModulons and significant activities during a hypoxia time course study found within our datasets (Peterson et al., 2019). During this study, the organism was exposed to changing dissolved oxygen levels, and we categorized the changes into 4 distinct phases. The time course starts with the Decreasing Oxygen phase, where dissolved oxygen levels start high and are lowered close to zero. This is followed by the Hypoxia Onset, where dissolved oxygen levels finally transition to zero. The 0% oxygen level is maintained all throughout the Stable Hypoxia phase, before finally increasing again in the Re-aeration phase. (Figure 6.1 a) Using this study, we first examined if iModulons reflected previously established models of *M. tuberculosis*'s TRN response to hypoxia. We then examined the activities of the iModulons in a phase specific manner across three of the four phases. DIMA plots were created comparing the iModulon activities from the first and last time point of each phase, and the significant iModulons were examined. We chose not to analyze the iModulons during stable hypoxia given the limited data.

The transcriptional changes associated with hypoxia are relatively well-characterized in *M. Tuberculosis*, and one of the benchmark studies for this response was performed by the Galagan group (Galagan et al., 2013). The study proposed a model of the organism's transcriptional network and determined that the DosR and Rv0081 TFs serve as the primary regulators for the hypoxic response while other TFs such as Rv2034, Rv3249c, KstR, and PhoP can alter the response. In order to confirm that iModulons reflect the known prior model, we examined iModulons mapped to hypoxia associated transcriptional factors and examined their activities throughout the hypoxic time course study. It is important to note that a few of the iModulons selected for analysis did not have direct one-to-one comparisons to known TFs. For example, there is no iModulon that fully captures the known regulon of Rv2034, so the Rv0078+Rv2034 iModulon was selected instead. Upon examining the six iModulons, we

found that the DevR (DosR), PhoP, KstR2, and Lsr2 iModulons had increased activity during the hypoxia time course. The two DevR iModulons showed the highest activity during the Hypoxia Onset phase, which confirms previous understanding that the DosR/DevR TF controls the hypoxia onset response (Figure 6.1 a) (Saini et al., 2004). Additionally, the increase in activity of the Lsr2, KstR2, and PhoP iModulons also further confirmed that iModulons capture the known transcriptional changes associated with hypoxia. While the Galagan paper cites the KstR transcriptional factor as being closely related to hypoxia response, no KstR iModulon was found within our ICA data structure. However, both KstR and KstR2 are described to have a role in cholesterol catabolism, and the presence suggests that KstR2 may play the same role as KstR during hypoxic response (Kendall et al., 2010). The Rv0078+Rv2034 and MbcA+Rv3249c+Rv3066 iModulons were not significantly expressed at any point in the time course.

3.7. Differing Levels of Oxygen Lead to Distinct Transcriptional States

The Decreasing Oxygen phase is the first phase of the time course, and represents the time when dissolved oxygen levels transition from 81% to 11%. Examining the most significantly changed iModulons during the Decreasing Oxygen phase, we discover a 3 part response in the bacterium. The first type of response is the significant increase in the production of enzymes associated with central carbon metabolism and energy production, and is captured by the Central Carbon Metabolism and Fumarate Reductase iModulons. Interestingly, the VirS iModulon, which is primarily associated with energy production and lipid metabolism, was found to have significantly decreased activity during this phase. The next type of response we found was increased activity in cell replication systems, which was captured by the upregulation of the Rv1828/SigH, GroEL-ES complex, and WhiB1 iModulons. Rv1828/SigH codes for a wide range of proteins with distinct functions, from lipoproteins to oxidases. However, we found that there were several genes found within the iModulon that had functions associated with DNA manipulation. These included cell division proteins (SepF, FtsZ), DNA helicases (RuvA/B/C), and DNA polymerases (Cole et al., 1998). Given the function of these genes, we concluded that the Rv1828/SigH is primarily involved in cell division. Both the WhiB1 and GroEL/ES complex iModulons are involved primarily with protein synthesis, as WhiB1 primarily contains ribosomal protein genes and the GroEL/ES complex contains multiple molecular chaperons. Additionally, WhiB1 contains several genes that code for RNA polymerase, which are essential for transcription. The presence of these iModulons and their associated genes suggests that up-regulating cell division genes is an important response in *M. tuberculosis* in a decreasing oxygen environment.

The final response of the Decreasing Oxygen phase was a shift in the mammalian cell entry (Mce) proteins produced within the cell. This response is captured by increased activity in the Mce1R iModulon and a decrease in activity for the Mce3R iModulon. The Mce proteins are invasive/adhesive cell surface proteins that promote virulence in *M. tuberculosis*, and play a role in invasion of host cells (Casali et al., 2006; Chitale et al., 2001; Singh et al., 2016). While the arrangement of each of these operons are remarkably similar, the expression of these operons have been shown to differ depending on the cell state, which is validated by the iModulon activity. Further examination of the Mce1R and Mce3R iModulons indicates that as the time course proceeds and the cell enters Hypoxia Onset and Stable

Hypoxia, the activities of the two iModulons reverse; the activity of Mce3R significantly increases while the activity of Mce1R significantly decreases. Given the close relation between hypoxia and virulence, we theorize that Mce1 proteins are important during the initial stages of infections and increased production of the proteins is triggered by decreasing oxygen levels. However, as hypoxia persists and the cells enter a dormant state, Mce3 proteins become more important for the survival of the cell, explaining the increased transcriptional activity in those genes.

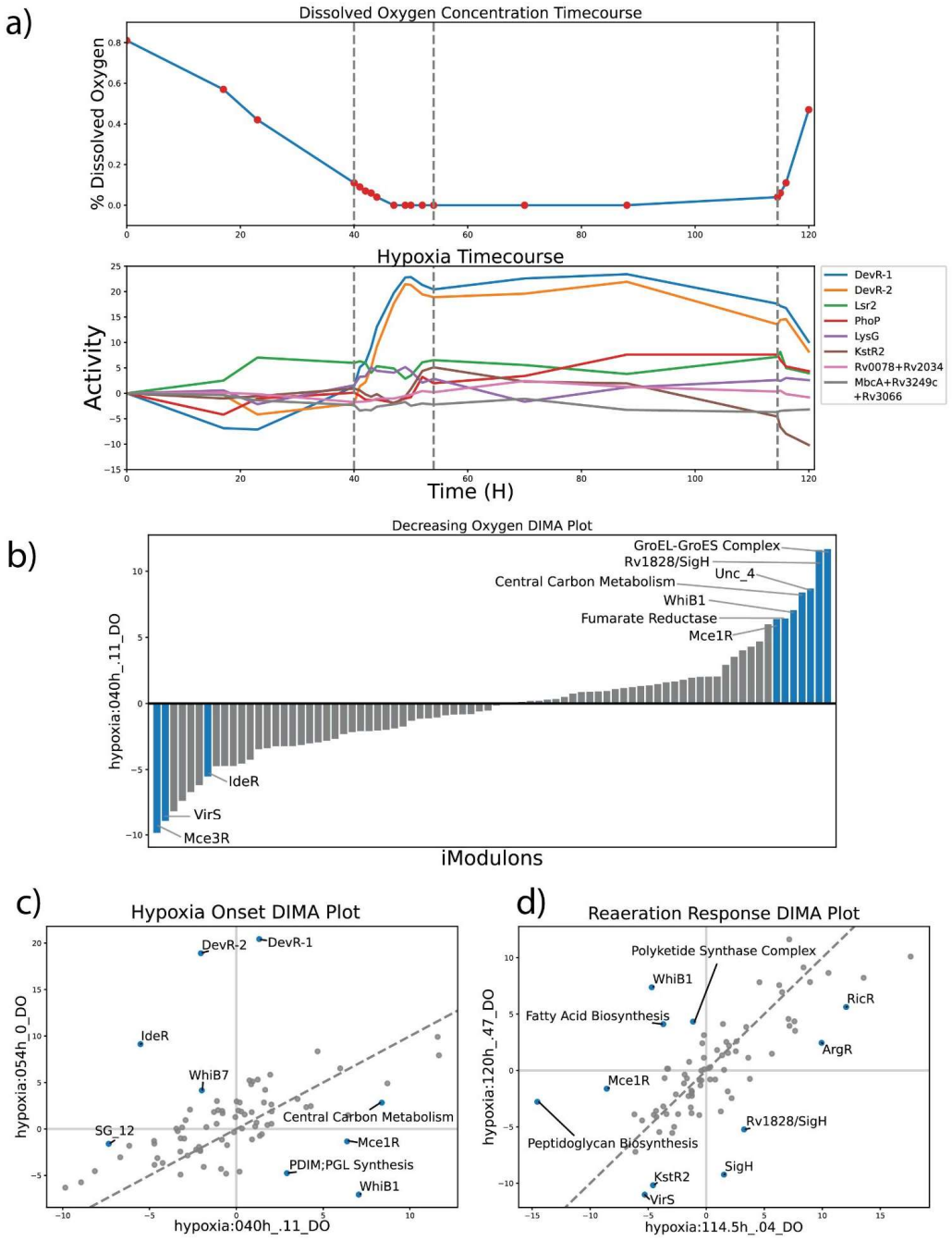

Figure 6.1: iModulons help Categorize the Phases of Hypoxia Response, including Metabolic Anticipation
(a) Time Course of *M. Tuberculosis* undergoing decreasing oxygen, hypoxia, and re-aeration. The top plot displays the dissolved oxygen concentration in the environment, and the bottom plot displays the activities over time for iModulons controlled by TFs previously identified to be highly involved in hypoxic response. (Galagan et al., 2013) The TF Rv2034 is represented by the iModulon Rv0078+Rv2034 and Rv3249c is represented by MbcA+Rv3249c+Rv3066 iModulons. (b) DIMA plots of hypoxia phases were created by comparing the iModulon activities between the first and last time point of each phase. The bar graph represents a 1D DIMA plot for the decreasing oxygen phase, since the original t=0 timepoint served as the reference condition. (c) DIMA plot for the Hypoxia Onset Phase (d) DIMA plot for the Reaeration phase.

The next phase of the hypoxia time course was the Hypoxia Onset phase, where the dissolved oxygen levels move from 11% to 0%. Apart from the previously described activities of both DevR iModoulons, we also found an interesting trend among the other iModulons with significant changes in activity. We found that a few of the iModulons had inverted activities during Hypoxia Onset when compared to the Decreasing Oxygen phase. The Mce1R, WhiB1, and Central Carbon Metabolism iModulons showed decreased activity over the course of the Hypoxia Onset phase, indicating that high activity in these systems are either unhelpful or even detrimental to the organism during complete hypoxia. On the other hand, the IdeR iModulon moved from decreasing in activity in the previous phase to significantly increasing in activity during Hypoxia Onset. Additionally, we found two new iModulons, the WhiB7 and PDIM;PGL Synthesis iModulons, with significant changes in activity during this phase. WhiB7 is a redox homeostasis transcriptional regulator that has been found to play a role in drug resistance (Burian et al., 2012). The PDIM;PGL Synthesis iModulon captures genes associated with the production of phthiocerol dimycocerosate (PDIM) and phenolic glycolipids (PGL). These family of molecules have been associated with cell wall impermeability, phagocytosis, defense against nitrosative and oxidative stress and, supposedly, biofilm formation (Ramos et al., 2020). The presence of both these systems during hypoxia is expected, though it is surprising that PDIM;PGL Synthesis was found to have decreased activity during Hypoxia Onset. This would suggest that though PDIM and PGL molecules may be important for oxidative stress defense, their production in a completely anaerobic environment may be detrimental to the survival of the cell.

The final phase of the hypoxia time course was the Re-aeration phase. During this phase, the cell returns to an aerobic environment as dissolved oxygen levels increase from 0% to 47%. Within this phase, we see significant change in several new iModulons. Most interesting among these iModulons are the Peptidoglycan Biosynthesis and Polyketide Synthase Complex. Both polyketides and peptidoglycans are cell membrane bound molecules that play a role in cell virulence and persistence. Peptidoglycans have been shown to be involved in cell growth and host response manipulation, while polyketides have been found to be essential in the formation of biofilms and are likely to improve *M. tuberculosis* persistence (Maitra et al., 2019; Pang et al., 2012). The increased activation of these iModulons under Reaeration suggests that *M. tuberculosis* is attempting to defend itself from a possible host response

during this phase. During this phase, we also found that the Fatty Acid Biosynthesis pathway had increased activation during this phase. We have previously described how KstR2, which is also involved in lipid and cholesterol metabolism, was also activated during this phase. Thus, it is quite clear to us that under reaeration conditions, *M. tuberculosis* moves from consumption of lipids and cholesterol to production. Taken all together, the hypoxia time course and iModulons allow us to describe the complex transcriptional response that *M. tuberculosis* undergoes throughout large shifts in oxygen concentration.

3.8. M. tuberculosis has Cell-Specific Transcriptional Responses Dependent on Host Cell Type

Because of the pathological impact of *M. tuberculosis*, we also examined two datasets which investigate the transcriptional response of *M. tuberculosis* during infection of host cells. In one dataset, *M. tuberculosis* was grown *in vitro* during infection of mice bone marrow derived macrophages (BMDM), and RNA-Seq was performed at 2, 8 and 24 hours after infection (Peterson et al., 2019). In the other dataset, *M. tuberculosis* was grown *in vivo* in mice neutrophils, and RNA-Seq was performed at a single time point after infection. (Grigorov A, Kondratieva T, Majorov K, Azhikina T, Apt AS, 2019) DIMA plots were created comparing each infection condition to a control at the same time point, and the significant iModulons were displayed (Figure 7.1). We also noted that many of the significantly changed iModulons between infectious and noninfectious conditions were similar to those found during hypoxia.

Examination of the significant iModulons under the three time points of the mice BMDM conditions found some consistent patterns. Across all time points, the acid sensing MarR iModulon was found to have decreased activity. MarR is a transcriptional repressor that allows *M. tuberculosis* to adapt to intracellular environments (Healy et al., 2016). In addition , we found that PrpR and Lipid Synthesis iModulons, along with the metal sensing Zur, M-box, and IdeR iModulons, were consistently downregulated throughout the infection time course. All of these iModulons have been shown to play a role in either starvation or hypoxia response, indicating that the residence within a macrophage requires distinct adaptations to multiple stresses (K. B. Arnvig et al., 2011; K. Arnvig & Young, 2012).

We also found changes in iModulons that did not appear at all timepoints, but are still informative about *M. tuberculosis*' response during infection. Early in the time course, *M. tuberculosis* showed increased activity in iModulons associated with central carbon metabolism and molecular export while decreasing activity in lipid and protein synthesis systems. Interestingly, the Peptidoglycan Biosynthesis and Leucine Related iModulons were also upregulated, and given the function of peptidoglycans this is a protective response (Maitra et al., 2019). However, as time proceeds we found that the many iModulons were down regulated, and eventually the organism reaches a dormant state by the 24 hour mark.

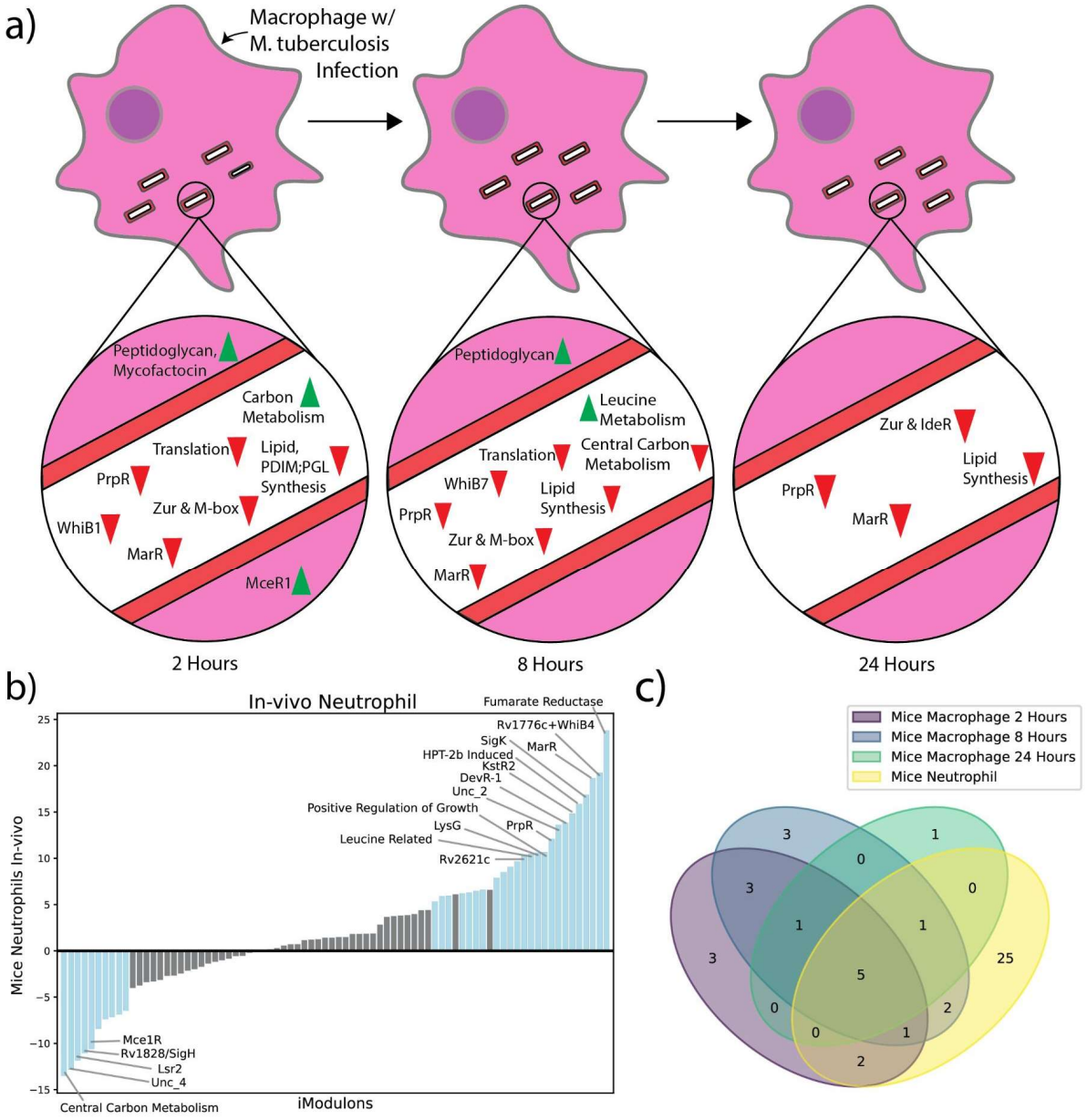

Figure 7.1: iModulon Response to Infection of Mice Macrophages and Neutrophils (a) A time course of the iModulon activities during infection of mice bone marrow derived macrophages (BMDM). The iModulons with differential activities at each time point are displayed as upregulated (green) or down regulated (red). Peptidoglycan, Mycofactocin, and MceR1 are displayed outside of the cell to indicate regulation of secretory pathways (b) 1D DIMA plot of differential iModulons between control non-infectious condition and in-vivo infection condition. Surprisingly, the most upregulated and most downregulated iModulons both regulate different portions of central carbon metabolism, which suggests that central carbon metabolism plays a large role in infection. (c) A core virulence response was constructed by examining the iModulons with differential activity across all virulence conditions (3 timepoints in mice macrophage infection and 1 neutrophil condition). The core virulence response was found to consist of KstR2, MarR, PrpR, Rv0681, Uncharacterized 2, and Zur.

3.9. Clustering of iModulon Activities Across All Conditions Reveal Consistent Stress Response

Through investigation of the iModulons across various conditions, we began to notice that certain sets of iModulons appeared to activate together. To investigate whether or not iModulons have similar activities to one another, we applied a method to calculate robust iModulon clusters based on the activities [cite anand when published], resulting in several clusters with biologically relevant implications. One such cluster is Cluster 36, which contains the DevR-1, DevR-2, and LysG iModulons (Figure 8.1) Given the function of the DevR TF and the presence of Rv0081 in LysG, it is clear that this cluster of iModulons captures the main hypoxic response in *M. tuberculosis*. (Galagan et al., 2013)

Clusters can also describe global responses in the *M. tuberculosis* TRN, as shown by Cluster 1 (Figure 8.1). Cluster 1 is one of the largest clusters obtained from the activity clustering method and contains a wide diversity of iModulons. These include virulence iModulons such as Mce1R, metal related iModulons such as RicR, and lipid metabolism iModulons such as Rv0681. Interestingly, we found that while 6 of the iModulons within the cluster were positively correlated with each other, Mce1R was found to be negatively correlated with the others. To help visualize which systems were controlled by this cluster, we mapped the genes within each cluster to known pathways using the iEK1008 COBRA model of *M. tuberculosis* (Kavvas et al., 2018). Upon mapping the iModulon genes to a metabolic map, we discovered that this cluster fully describes pathways related to cholesterol catabolism into propionyl-CoA. As mentioned previously, propionyl-CoA is an important precursor to sulfolipids, and we found that Cluster 1 does control pathways associated with sulfur import , activation of sulfur, and the formation of sulfolipids. We also found that the cluster controls the production of mce1 proteins, the type 1 NADH-dehydrogenase, and metal sensing systems. Type 1 NADH-dehydrogenase is known to produce ROS species and increase oxidative stress, while metal sensing systems such as those coded by RicR are important for protection against oxidative stress (Larosa & Remacle, 2018; Ward et al., 2010). Given the function of the genes found within the cluster, we propose that Cluster 1 represents a general stress response in *M. tuberculosis*, most likely related to intra-host survival. While the true function of Cluster 1 has yet to be confirmed, we find that the clustering of iModulons can allow us to gain a complex network level understanding of TF response.

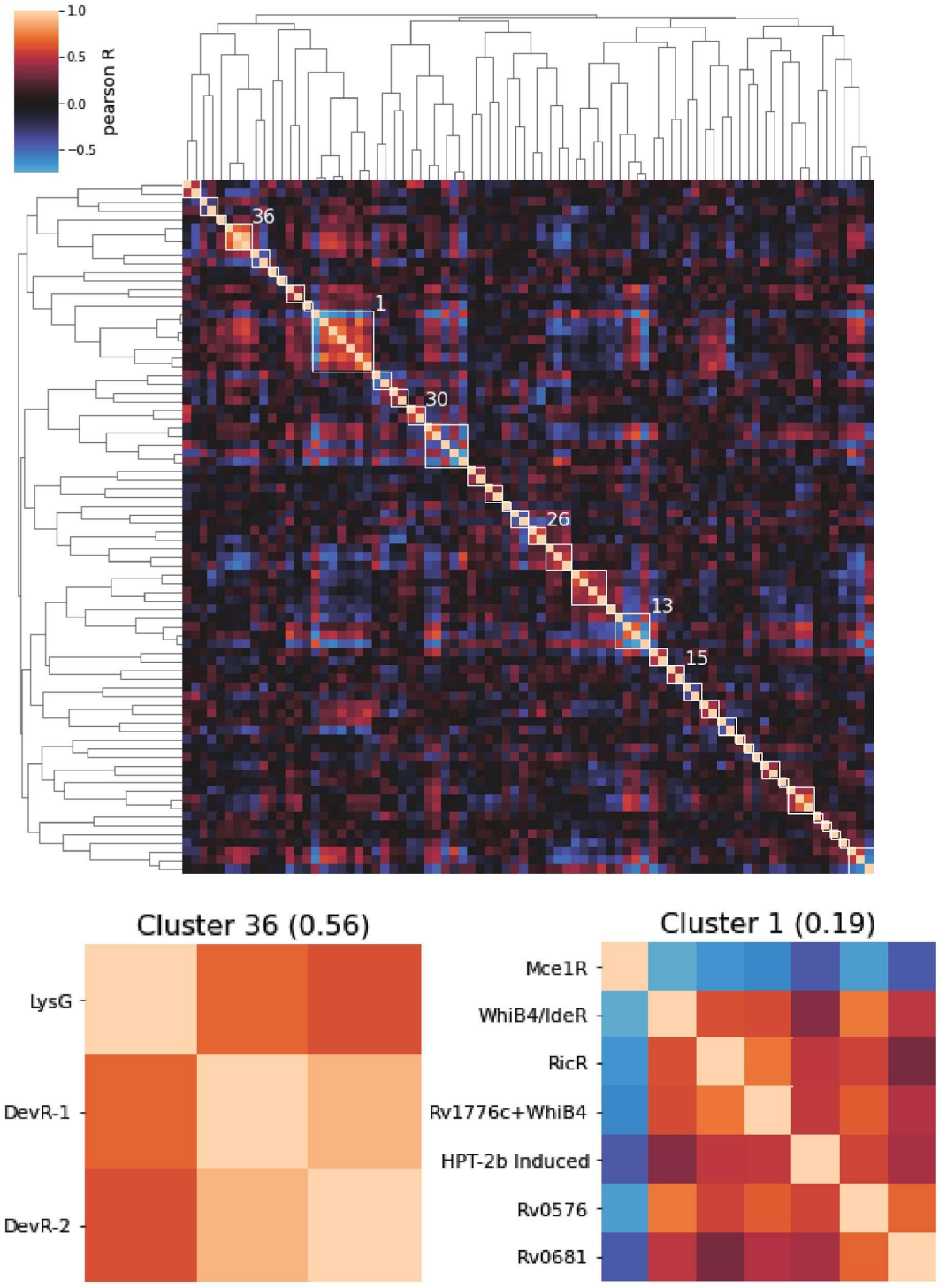

Figure 8.1: Clustermap of M. Tuberculosis Activities
Pairwise PearsonR correlations were computed and iModulons with similar activities across conditions were clustered together via agglomerative clustering. Cluster 36 confirms the relationship between DevR-1 and DevR-2, and illuminates the relationship of LysG with the two DevR iModulons. Given that the Rv0081 TF, which regulates the main hypoxic response in M. tuberculosis along with DevR/DosR, is present within the LysG iModulon, we believe that this is a valid relationship.(Galagan et al., 2013) Cluster 1 describes a possible global stress response in M. tuberculosis that spans 6 different iModulons. Given the activation of many of these iModulons in under lipid, redox, and virulence conditions, we suggest that this is likely related to infection of host cells.

Chapter 4: Discussion

4.1: Discussion

Here, we utilized ICA to decompose 657 separate RNA-seq profiles of M. tuberculosis. We extracted 80 independently modulating iModulons, many of the corresponding to important transcription factors within the organism. Analysis of iModulons mapped to known transcription factors reveals that the machine learning technique maintains the structure and biological function of the known regulon. Additionally, mapping of iModulons to previously predicted regulons and analysis of the cluster's activity revealed the function of several new transcription factors. 9 iModulons remain to be mapped to known regulons, but exploration of their functions may also reveal new transcription factors.

System-based analysis of iModulon activity across the various environmental conditions both validates well characterized transcription factor behaviour and reveals novel TRN responses. iModulons helped validate the upregulation of the glyoxylate shunt during metabolism of lactate and pyruvate, as well as revealed the activation of sulfur and propionyl-CoA metabolism under the same conditions. iModulons helped redefine the core lipid response in M. tuberculosis, and helped elucidate the relationship between oxygen levels, time, and the TRN. ICA decomposition also revealed that M. tuberculosis has a distinct transcriptional response during infection depending on the type of host cell. Finally, clustering by activity reveals a novel global stress response that is composed of 7 different iModulons, a response that is likely key to virulence in M. tuberculosis.

The iModulons composition and activities are available on Github (***), and researchers interested in examining the data in greater detail can find a pre-built website in the repository (via html files). Additionally, the code used in the analysis of the data and the creation of the figures used in this paper can be found in the repository as well. Given the size of the dataset, this data still has potential to reveal new insights into the function of uncharacterized transcription factors and the TRN behaviour of M.tuberculosis under different conditions.

Future research utilizing ICA decomposition of M. tuberculosis and other mycobacterium still holds great potential. One limitation of this paper was a lack of analysis of M. tuberculosis under antibiotic conditions. Given the rise of multidrug resistant and extensively drug resistant strains, understanding the TRN behaviour of M. tuberculosis under antibiotic conditions may give additional insights on methods to

combat the pathogen. (World Health Organization, 2020) Datasets such as the Tuberculosis Antibiotic Resistance Catalog project are gold mines of data waiting for further analysis via ICA. (Velayati et al., 2016; Winglee et al., 2016) Additionally, analysis of different strains used to study H37Rv are also great targets for ICA decomposition, as such analysis can reveal strain specific differences in bacterial transcription. In summary, the ICA decomposition on mycobacterium transcription data is still rich for new discoveries.

References

1. Aguilar-Ayala, D. A., Tilleman, L., Van Nieuwerburgh, F., Deforce, D., Palomino, J. C., Vandamme, P., Gonzalez-Y-Merchand, J. A., & Martin, A. (2017). The transcriptome of Mycobacterium tuberculosis in a lipid-rich dormancy model through RNAseq analysis. *Scientific Reports*, *7*(1), 17665.

2. Ahidjo, B. A., Kuhnert, D., McKenzie, J. L., Machowski, E. E., Gordhan, B. G., Arcus, V., Abrahams, G. L., & Mizrahi, V. (2011). VapC toxins from Mycobacterium tuberculosis are ribonucleases that differentially inhibit growth and are neutralized by cognate VapB antitoxins. *PloS One*, *6*(6), e21738.

3. Arnvig, K. B., Comas, I., Thomson, N. R., Houghton, J., Boshoff, H. I., Croucher, N. J., Rose, G., Perkins, T. T., Parkhill, J., Dougan, G., & Young, D. B. (2011). Sequence-based analysis uncovers an abundance of non-coding RNA in the total transcriptome of Mycobacterium tuberculosis. *PLoS Pathogens*, *7*(11), e1002342.

4. Arnvig, K., & Young, D. (2012). Non-coding RNA and its potential role in Mycobacterium tuberculosis pathogenesis. *RNA Biology*, *9*(4), 427–436.

5. Bansal, R., Anil Kumar, V., Sevalkar, R. R., Singh, P. R., & Sarkar, D. (2017). Mycobacterium tuberculosis virulence-regulator PhoP interacts with alternative sigma factor SigE during acid-stress response. *Molecular Microbiology*, *104*(3), 400–411.

6. Bartek, I. L., Woolhiser, L. K., Baughn, A. D., Basaraba, R. J., Jacobs, W. R., Jr, Lenaerts, A. J., & Voskuil, M. I. (2014). Mycobacterium tuberculosis Lsr2 is a global transcriptional regulator required for adaptation to changing oxygen levels and virulence. *mBio*, *5*(3), e01106–e01114.

7. Boot, M., Commandeur, S., Subudhi, A. K., Bahira, M., Smith, T. C., 2nd, Abdallah, A. M., van Gemert, M., Lelièvre, J., Ballell, L., Aldridge, B. B., Pain, A., Speer, A., & Bitter, W. (2018). Accelerating Early Antituberculosis Drug Discovery by Creating Mycobacterial Indicator Strains That Predict Mode of Action. *Antimicrobial Agents and Chemotherapy*, *62*(7). https://doi.org/10.1128/AAC.00083-18

8. Briffotaux, J., Huang, W., Wang, X., & Gicquel, B. (2017). MmpS5/MmpL5 as an efflux pump in Mycobacterium species. *Tuberculosis* , *107*, 13–19.

9. Burian, J., Ramón-García, S., Sweet, G., Gómez-Velasco, A., Av-Gay, Y., & Thompson, C. J. (2012). The mycobacterial transcriptional regulator whiB7 gene links redox homeostasis and intrinsic antibiotic resistance. *The Journal of Biological Chemistry*, *287*(1), 299–310.

10. Casali, N., White, A. M., & Riley, L. W. (2006). Regulation of the Mycobacterium tuberculosis mce1 operon. *Journal of Bacteriology*, *188*(2), 441–449.

11. Cellier, M. F. M. (2012). Nramp: from sequence to structure and mechanism of divalent metal import. *Current Topics in Membranes*, *69*, 249–293.

12. Chitale, S., Ehrt, S., Kawamura, I., Fujimura, T., Shimono, N., Anand, N., Lu, S., Cohen-Gould, L., & Riley, L. W. (2001). Recombinant Mycobacterium tuberculosis protein associated with mammalian cell entry. *Cellular Microbiology*, *3*(4), 247–254.

13. Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S. V., Eiglmeier, K., Gas, S., Barry, C. E., 3rd, Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., … Barrell, B. G. (1998). Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature*, *393*(6685), 537–544.

14. Delogu, G., Sali, M., & Fadda, G. (2013). The biology of mycobacterium tuberculosis infection.

Mediterranean Journal of Hematology and Infectious Diseases, *5*(1), e2013070.

15. Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, *35*(4), 316–319.

16. Du, P., Sohaskey, C. D., & Shi, L. (2016). Transcriptional and Physiological Changes during Mycobacterium tuberculosis Reactivation from Non-replicating Persistence. *Frontiers in Microbiology*, *7*, 1346.

17. Ehrt, S., & Schnappinger, D. (2007). Mycobacterium tuberculosis virulence: lipids inside and out. *Nature Medicine*, *13*(3), 284–285.

18. Eroshenko, D. V., Polyudova, T. V., & Pyankova, A. A. (2020). VapBC and MazEF toxin/antitoxin systems in the regulation of biofilm formation and antibiotic tolerance in nontuberculous mycobacteria. *International Journal of Mycobacteriology*, *9*(2), 156–166.

19. Ester, M., Kriegel, H.-P., Sander, J., Xu, X., & Others. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Kdd*, *96*, 226–231.

20. Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* , *32*(19), 3047–3048.

21. Feng, L., Chen, S., & Hu, Y. (2018). PhoPR Positively Regulates whiB3 Expression in Response to Low pH in Pathogenic Mycobacteria. *Journal of Bacteriology*, *200*(8). https://doi.org/10.1128/JB.00766-17

22. Forrellad, M. A., McNeil, M., Santangelo, M. de la P., Blanco, F. C., García, E., Klepp, L. I., Huff, J., Niederweis, M., Jackson, M., & Bigi, F. (2014). Role of the Mce1 transporter in the lipid homeostasis of Mycobacterium tuberculosis. *Tuberculosis* , *94*(2), 170–177.

23. Galagan, J. E., Minch, K., Peterson, M., Lyubetskaya, A., Azizi, E., Sweet, L., Gomes, A., Rustad, T., Dolganov, G., Glotova, I., Abeel, T., Mahwinney, C., Kennedy, A. D., Allard, R., Brabant, W., Krueger, A., Jaini, S., Honda, B., Yu, W.-H., … Schoolnik, G. K. (2013). The Mycobacterium tuberculosis regulatory network and hypoxia. *Nature*, *499*(7457), 178–183.

24. Gonzalo-Asensio, J., Mostowy, S., Harders-Westerveen, J., Huygen, K., Hernández-Pando, R., Thole, J., Behr, M., Gicquel, B., & Martín, C. (2008). PhoP: a missing piece in the intricate puzzle of Mycobacterium tuberculosis virulence. *PloS One*, *3*(10), e3496.

25. Gordon, B. R. G., Imperial, R., Wang, L., Navarre, W. W., & Liu, J. (2008). Lsr2 of Mycobacterium represents a novel class of H-NS-like proteins. *Journal of Bacteriology*, *190*(21), 7052–7059.

26. Grigorov A, Kondratieva T, Majorov K, Azhikina T, Apt AS. (2019). *Transcriptional response of Mycobacterium tuberculosis in mouse peritoneal neutrophils* [Data set]. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE140156

27. Healy, C., Golby, P., MacHugh, D. E., & Gordon, S. V. (2016). The MarR family transcription factor Rv1404 coordinates adaptation of Mycobacterium tuberculosis to acid stress via controlled expression of Rv1405c, a virulence-associated methyltransferase. *Tuberculosis* , *97*, 154–162.

28. Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., von Mering, C., & Bork, P. (2017). Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Molecular Biology and Evolution*, *34*(8), 2115–2122.

29. Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks / a Publication of the IEEE Neural Networks Council*, *10*(3), 626–634.

30. Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M., & Tanabe, M. (2021). KEGG: integrating viruses and cellular organisms. *Nucleic Acids Research*, *49*(D1), D545–D551.

31. Kans, J. (2020). Entrez direct: E-utilities on the UNIX command line. In *Entrez Programming Utilities Help [Internet]*. National Center for Biotechnology Information (US).

32. Karp, P. D., Billington, R., Caspi, R., Fulcher, C. A., Latendresse, M., Kothari, A., Keseler, I. M., Krummenacker, M., Midford, P. E., Ong, Q., Ong, W. K., Paley, S. M., & Subhraveti, P. (2019). The BioCyc collection of microbial genomes and metabolic pathways. *Briefings in Bioinformatics*, *20*(4), 1085–1093.

33. Kavvas, E. S., Seif, Y., Yurkovich, J. T., Norsigian, C., Poudel, S., Greenwald, W. W., Ghatak, S., Palsson, B. O., & Monk, J. M. (2018). Updated and standardized genome-scale reconstruction of Mycobacterium tuberculosis H37Rv, iEK1011, simulates flux states indicative of physiological conditions. *BMC Systems Biology*, *12*(1), 25.

34. Kelkar, D. S., Kumar, D., Kumar, P., Balakrishnan, L., Muthusamy, B., Yadav, A. K., Shrivastava, P., Marimuthu, A., Anand, S., Sundaram, H., Kingsbury, R., Harsha, H. C., Nair, B., Prasad, T. S. K., Chauhan, D. S., Katoch, K., Katoch, V. M., Kumar, P., Chaerkady, R., … Pandey, A. (2011). Proteogenomic analysis of Mycobacterium tuberculosis by high resolution mass spectrometry. *Molecular & Cellular Proteomics: MCP*, *10*(12), M111.011627.

35. Kendall, S. L., Burgess, P., Balhana, R., Withers, M., Ten Bokum, A., Lott, J. S., Gao, C., Uhia-Castro, I., & Stoker, N. G. (2010). Cholesterol utilization in mycobacteria is controlled by two TetR-type transcriptional regulators: kstR and kstR2. *Microbiology*, *156*(Pt 5), 1362–1371.

36. Kodama, Y., Shumway, M., Leinonen, R., & International Nucleotide Sequence Database Collaboration. (2012). The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Research*, *40*(Database issue), D54–D56.

37. Kołodziej, M., Trojanowski, D., Bury, K., Hołówka, J., Matysik, W., Kąkolewska, H., Feddersen, H., Giacomelli, G., Konieczny, I., Bramkamp, M., & Zakrzewska-Czerwińska, J. (2021). Lsr2, a nucleoid-associated protein influencing mycobacterial cell cycle. *Scientific Reports*, *11*(1), 2910.

38. Kumar, P., Schelle, M. W., Jain, M., Lin, F. L., Petzold, C. J., Leavell, M. D., Leary, J. A., Cox, J. S., & Bertozzi, C. R. (2007). PapA1 and PapA2 are acyltransferases essential for the biosynthesis of the Mycobacterium tuberculosis virulence factor sulfolipid-1. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(27), 11221–11226.

39. Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*(3), R25.

40. Larosa, V., & Remacle, C. (2018). Insights into the respiratory chain and oxidative stress. *Bioscience Reports*, *38*(5). https://doi.org/10.1042/BSR20171492

41. Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* , *30*(7), 923–930.

42. Ly, A., & Liu, J. (2020). Mycobacterial Virulence Factors: Surface-Exposed Lipids and Secreted Proteins. *International Journal of Molecular Sciences*, *21*(11). https://doi.org/10.3390/ijms21113985

43. Maciąg, A., Dainese, E., Marcela Rodriguez, G., Milano, A., Provvedi, R., Pasca, M. R., Smith, I., Palù, G., Riccardi, G., & Manganelli, R. (2007). Global Analysis of the Mycobacterium tuberculosis Zur (FurB) Regulon. *Journal of Bacteriology*, *189*(3), 730–740.

44. Maitra, A., Munshi, T., Healy, J., Martin, L. T., Vollmer, W., Keep, N. H., & Bhakta, S. (2019). Cell wall peptidoglycan in Mycobacterium tuberculosis: An Achilles' heel for the TB-causing pathogen.

FEMS Microbiology Reviews, *43*(5), 548–575.

45. Ma, S., Jaipalli, S., Larkins-Ford, J., Lohmiller, J., Aldridge, B. B., Sherman, D. R., & Chandrasekaran, S. (2019). Transcriptomic Signatures Predict Regulators of Drug Synergy and Clinical Regimen Efficacy against Tuberculosis. *mBio*, *10*(6). https://doi.org/10.1128/mBio.02627-19

46. Mishra, R., Kohli, S., Malhotra, N., Bandyopadhyay, P., Mehta, M., Munshi, M., Adiga, V., Ahuja, V. K., Shandil, R. K., Rajmani, R. S., Seshasayee, A. S. N., & Singh, A. (2019). Targeting redox heterogeneity to counteract drug tolerance in replicating Mycobacterium tuberculosis. *Science Translational Medicine*, *11*(518). https://doi.org/10.1126/scitranslmed.aaw6635

47. Neyrolles, O., Wolschendorf, F., Mitra, A., & Niederweis, M. (2015). Mycobacteria, metals, and the macrophage. *Immunological Reviews*, *264*(1), 249–263.

48. Pandey, R., Russo, R., Ghanny, S., Huang, X., Helmann, J., & Rodriguez, G. M. (2015). MntR(Rv2788): a transcriptional regulator that controls manganese homeostasis in Mycobacterium tuberculosis. *Molecular Microbiology*, *98*(6), 1168–1183.

49. Pang, J. M., Layre, E., Sweet, L., Sherrid, A., Moody, D. B., Ojha, A., & Sherman, D. R. (2012). The polyketide Pks1 contributes to biofilm formation in Mycobacterium tuberculosis. *Journal of Bacteriology*, *194*(3), 715–721.

50. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research: JMLR*, *12*(Oct), 2825–2830.

51. Pérez, E., Samper, S., Bordas, Y., Guilhot, C., Gicquel, B., & Martín, C. (2001). An essential role for phoP in Mycobacterium tuberculosis virulence. *Molecular Microbiology*, *41*(1), 179–187.

52. Peterson, E. J., Bailo, R., Rothchild, A. C., Arrieta-Ortiz, M. L., Kaur, A., Pan, M., Mai, D., Abidi, A. A., Cooper, C., Aderem, A., Bhatt, A., & Baliga, N. S. (2019). Path-seq identifies an essential mycolate remodeling program for mycobacterial host adaptation. *Molecular Systems Biology*, *15*(3), e8584.

53. Poudel, S., Tsunemoto, H., Seif, Y., Sastry, A. V., Szubin, R., Xu, S., Machado, H., Olson, C. A., Anand, A., Pogliano, J., Nizet, V., & Palsson, B. O. (2020). Revealing 29 sets of independently modulated genes in Staphylococcus aureus, their regulators, and role in key physiological response. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(29), 17228–17239.

54. Ramos, B., Gordon, S. V., & Cunha, M. V. (2020). Revisiting the expression signature of pks15/1 unveils regulatory patterns controlling phenolphtiocerol and phenolglycolipid production in pathogenic mycobacteria. In *bioRxiv* (p. 2020.02.20.950329). https://doi.org/10.1101/2020.02.20.950329

55. Rodriguez, G. M., Voskuil, M. I., Gold, B., Schoolnik, G. K., & Smith, I. (2002). ideR, An essential gene in mycobacterium tuberculosis: role of IdeR in iron-dependent gene expression, iron metabolism, and oxidative stress response. *Infection and Immunity*, *70*(7), 3371–3381.

56. Rustad, T. R., Sherrid, A. M., Minch, K. J., & Sherman, D. R. (2009). Hypoxia: a window into Mycobacterium tuberculosis latency. *Cellular Microbiology*, *11*(8), 1151–1159.

57. Rychel, K., Sastry, A. V., & Palsson, B. O. (2020). Machine learning uncovers independently regulated modules in the Bacillus subtilis transcriptome. In *Cold Spring Harbor Laboratory* (p. 2020.04.26.062638). https://doi.org/10.1101/2020.04.26.062638

58. Saini, D. K., Malhotra, V., Dey, D., Pant, N., Das, T. K., & Tyagi, J. S. (2004). DevR-DevS is a bona fide two-component system of Mycobacterium tuberculosis that is hypoxia-responsive in the absence of the DNA-binding domain of DevR. *Microbiology*, *150*(Pt 4), 865–875.

59. Sala, C., Haouz, A., Saul, F. A., Miras, I., Rosenkrands, I., Alzari, P. M., & Cole, S. T. (2009). Genome-wide regulon and crystal structure of BlaI (Rv1846c) from Mycobacterium tuberculosis. *Molecular Microbiology*, *71*(5), 1102–1116.

60. Sastry, A. V., Gao, Y., Szubin, R., Hefner, Y., Xu, S., Kim, D., Choudhary, K. S., Yang, L., King, Z. A., & Palsson, B. O. (2019). The Escherichia coli transcriptome mostly consists of independently regulated modules. *Nature Communications*, *10*(1), 5536.

61. Sastry, A. V., Hu, A., Heckmann, D., Poudel, S., Kavvas, E., & Palsson, B. O. (2021). Independent component analysis recovers consistent regulatory signals from disparate datasets. *PLoS Computational Biology*, *17*(2), e1008647.

62. Serafini, A., Tan, L., Horswell, S., Howell, S., Greenwood, D. J., Hunt, D. M., Phan, M.-D., Schembri, M., Monteleone, M., Montague, C. R., Britton, W., Garza-Garcia, A., Snijders, A. P., VanderVen, B., Gutierrez, M. G., West, N. P., & de Carvalho, L. P. S. (2019). Mycobacterium tuberculosis requires glyoxylate shunt and reverse methylcitrate cycle for lactate and pyruvate metabolism. *Molecular Microbiology*, *112*(4), 1284–1307.

63. Singh, P., Katoch, V. M., Mohanty, K. K., & Chauhan, D. S. (2016). Analysis of expression profile of mce operon genes (mce1, mce2, mce3 operon) in different Mycobacterium tuberculosis isolates at different growth phases. *The Indian Journal of Medical Research*, *143*(4), 487–494.

64. Tang, S., Hicks, N. D., Cheng, Y.-S., Silva, A., Fortune, S. M., & Sacchettini, J. C. (2019). Structural and functional insight into the Mycobacterium tuberculosis protein PrpR reveals a novel type of transcription factor. *Nucleic Acids Research*, *47*(18), 9934–9949.

65. The Gene Ontology Consortium. (2019). The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, *47*(D1), D330–D338.

66. Tufariello, J. M., Chapman, J. R., Kerantzas, C. A., Wong, K.-W., Vilchèze, C., Jones, C. M., Cole, L. E., Tinaztepe, E., Thompson, V., Fenyö, D., Niederweis, M., Ueberheide, B., Philips, J. A., & Jacobs, W. R., Jr. (2016). Separable roles for Mycobacterium tuberculosis ESX-3 effectors in iron acquisition and virulence. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(3), E348–E357.

67. Turkarslan, S., Peterson, E. J. R., Rustad, T. R., Minch, K. J., Reiss, D. J., Morrison, R., Ma, S., Price, N. D., Sherman, D. R., & Baliga, N. S. (2015). A comprehensive map of genome-wide gene regulation in Mycobacterium tuberculosis. *Scientific Data*, *2*, 150010.

68. UniProt Consortium. (2021). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, *49*(D1), D480–D489.

69. Velayati, A. A., Abeel, T., Shea, T., Konstantinovich Zhavnerko, G., Birren, B., Cassell, G. H., Earl, A. M., Hoffner, S., & Farnia, P. (2016). Populations of latent Mycobacterium tuberculosis lack a cell wall: Isolation, visualization, and whole-genome characterization. *International Journal of Mycobacteriology*, *5*(1), 66–73.

70. Wang, L., Wang, S., & Li, W. (2012). RSeQC: quality control of RNA-seq experiments. *Bioinformatics* , *28*(16), 2184–2185.

71. Ward, S. K., Abomoelak, B., Hoye, E. A., Steinberg, H., & Talaat, A. M. (2010). CtpV: a putative copper exporter required for full virulence of Mycobacterium tuberculosis. *Molecular Microbiology*, *77*(5), 1096–1110.

72. Waskom, M., Botvinnik, O., Hobson, P., Warmenhoven, J., Cole, J. B., Halchenko, Y., Vanderplas, J., Hoyer, S., Villalba, S., Quintero, E., Miles, A., Augspurger, T., Yarkoni, T., Evans, C., Wehner, D., Rocher, L., Megies, T., Coelho, L. P., Ziegler, E., Hoppe T., Seabold S., Pascual S., Cloud P., Koskinen M., Hausler C., Jemmet K., Milajevs D., Qalieh A., Allan D., Meyer, K. (2015). Seaborn: V0.6.0 (June 2015). In *Zenodo*. https://doi.org/10.5281/zenodo.19108

73. Wilburn, K. M., Fieweger, R. A., & VanderVen, B. C. (2018). Cholesterol and fatty acids grease the wheels of Mycobacterium tuberculosis pathogenesis. *Pathogens and Disease*, *76*(2). https://doi.org/10.1093/femspd/fty021

74. Winglee, K., Manson McGuire, A., Maiga, M., Abeel, T., Shea, T., Desjardins, C. A., Diarra, B., Baya, B., Sanogo, M., Diallo, S., Earl, A. M., & Bishai, W. R. (2016). Whole Genome Sequencing of Mycobacterium africanum Strains from Mali Provides Insights into the Mechanisms of Geographic Restriction. *PLoS Neglected Tropical Diseases*, *10*(1), e0004332.

75. World Health Organization. (2020). *GLOBAL TUBERCULOSIS REPORT 2020*. https://www.who.int/tb/en/

76. Zheng, X., Papavinasasundaram, K. G., & Av-Gay, Y. (2007). Novel substrates of Mycobacterium tuberculosis PknH Ser/Thr kinase. *Biochemical and Biophysical Research Communications*, *355*(1), 162–168.

77. Ziemann, M., Kaspi, A., & El-Osta, A. (2019). Digital expression explorer 2: a repository of uniformly processed RNA sequencing data. *GigaScience*, *8*(4). https://doi.org/10.1093/gigascience/giz022