Patterns in Pathogenesis:
Elucidating bacterial host interaction

Niels A. Zondervan

Propositions

1. Pathogenesis is an emergent property of biological systems.
   (this thesis)

2. The only way to obtain reproducible data is to ensure already at acquisition that data remains findable, accessible, interoperable, and reusable.
   (this thesis)

3. Semantically annotating data enables rapid data exploration and research.

4. Mathematical models of metabolism are misleading by suggesting exactness.

5. Natural systems balance order and chaos at all scales.

6. For a rapid switch from a fossil fuel-based economy to a green economy, incentives are required.

7. It is advantageous to reduce the 40-hour workweek to 24-hours.

Propositions belonging to the thesis, entitled

Patterns in Pathogenesis: Elucidating bacterial host interaction

Niels A. Zondervan
Wageningen 5 December 2022

Patterns in Pathogenesis:

Elucidating bacterial host interaction

Niels A. Zondervan

Thesis Committee

Promotors
Prof. Dr VAP Martins Dos Santos
Professor Biomanufacturing & Digital Twins
Wageningen University & Research

Prof. Dr M Suarez Diez
Professor of Systems and Synthetic Biology
Wageningen University & Research

Other members
Prof. Dr AH Kersten, Wageningen University & Research
Prof. Dr S. Haussler, Helmholtz Institute for Infection Diseases, Germany
Dr L. Garcia Morales, MSD, Boxmeer, The Netherlands
Dr M. Svensson, Karolinska Institute, Stockholm, Sweden

This research was conducted under the auspices of the Graduate School VLAG (Advanced studies in Food Technology, Agrobiotechnology, Nutrition and Health Sciences).

Patterns in Pathogenesis:

Elucidating bacterial host interaction

Niels A. Zondervan

Thesis
submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Monday 5 December 2022
at 4 p.m. in the Omnia Auditorium.

(Niels) Nicolaas Adriaan Zondervan
Patterns in pathogenesis
195 pages

PhD thesis, Wageningen University, Wageningen, the Netherlands (2022)
With references, with summary in English

ISBN: 978-94-6447-446-6
DOI: https://doi.org/10.18174/578902

Lists of Abbreviations

AC	Adenylate cyclase
AcCoA	Acetyl coenzyme A
ACE	Acetate
ACK	Acetate kinase
ADP	Adenosine diphosphate
ATP	Adenosine triphosphate
ATPase	Adenylpyrophosphatase
cAMP	Adenylate cyclase Cyclic adenosine monophosphate
cAMP	Cyclic adenosine monophosphate
CMR	Cyclic-AMP and redox responsive transcription factor Cyclic-AMP dependent regulatory protein Diacylglycerol
CoA	Coenzyme A dehydrogenase
CRP	Cyclic-AMP dependent regulatory protein
DAG	Diacylglycerol
DevRST	DevRST is a two-component regulator and sensor, which regulate genes coding for proteins that help Mtb prepare for dormancy and subsequent resuscitation
DGP	Diacylglycerol phosphate
ENO	Enolase
EspR	A virulence associated transcriptional regulator upregulated by PhoP
F6P	Fructose 6-phosphate
FBA	Fructose-bisphosphate aldolase
FBP	Fructose-1,6-bisphosphatase
G6P	Glucose 6-phosphate
GAP	Glyceraldehyde-3-phosphate
GAPDH	Glyceraldehyde 3-phosphate
GLC Ext	Extracellular glucose
GMP	Glycerate phosphomutase
IdeR	Iron-dependent regulator
Kcat	Enzyme catalytic rate
Keq	Equilibrium constant
Km	Michaelis-Menten constant
KO	Knock out
LAC	Lactate
LDH	Lactate dehydrogenase
Lsr2	A histone like regulator that binds AT-rich regions virulence islands, acting as a global regulator to aid in the adaptation to extremes in oxygen availability
MntR	Manganese-dependent transcriptional repressor
MPN	Mycoplasma pneumonia
MprAB	A two component sensor and regulator that responds to cell envelop stress
Mtb	Mycobacterium tuberculosis
NAD	Nicotinamide adenine dinucleotide
NADH	Reduced nicotinamide adenine dinucleotide
OE	Over-expression
PDH	Pyruvate dehydrogenase
PDIM	Phthiocerol dimycocerosates

PEP	Phosphoenolpyruvate
PFK	Phosphofructokinase
PGI	Phosphoglucose isomerase
PGK	Phosphoglycerate kinase
PhoPR	Two component system, regulator and kinase which regulate many (virulence-)genes involved in adaptation to hypoxia (low oxygen).
Pi Int	Orthophosphate
PTA	Phosphotransacetylase
PTS Glc	Phosphotransferase system
PYK	Pyruvate kinase
PYR	Pyruvate
SigE	Extracytoplasmic alternative Sigma factor E, involved in response to low pH and cell stress
Zur	Zinc uptake regulator

Table of Contents

Chapter 1
Introduction

1.1 Bacterial pathogens – to be or not to be

Bacteria have adapted to a great many environments. Bacteria can live on icy glaciers, in hot underwater geysers or even on radioactive material in uranium mines [1], [2]. From the food we eat, to the water we drink and the air that we breathe, everywhere we find bacteria. Bacteria have evolved to live in nearly any environment and can break down nearly any chemical component we know of. It should therefore not come as a surprise that bacteria have also evolved to live together with higher organisms such as plants insects, animals, and humans. Some of these bacteria live in harmony with their host with either one or both organisms benefitting. We call these symbiotic bacteria *commensals*. Commensal bacteria enable, for instance, their host organisms to break down complex metabolites in their gut and help protect them against bacterial and viral pathogens [3]. However, not all bacteria live in harmony with their host. Bacteria that live on a host causing illness, we call pathogenic bacteria. However, there is no clear separation between "friendly" commensal bacteria and "unfriendly" pathogenic bacteria. Commensal bacteria, such as *Staphylococcus aureus,* can turn pathogenic when they get access to the bloodstream [4]. "Bad" bacteria, like *Streptococcus pneumoniae,* can in in some cases help protect against infection by other "bad" bacteria like *S. aureus* [5]. Whether bacteria contribute to the wellbeing of a host depends on a multitude of factors, such as a) the complex interplay with the microbial community present, b), the overall health and immune state of the host, c) abundance of available nutrients in the host environment d) the tissue or location at which bacteria occur in the host. All these factors in the system at large, determine the wellbeing of both the bacteria and the host they live on.

1.2 Model pathogenic bacteria

Within this thesis various pathogens were studied using common concepts, methods, and strategies. *M. tuberculosis, M. pneumoniae, Staphylococcus* and *Streptococcus* species such as *Staphylococcus aureus* and *Streptococcus pneumoniae* were studied. These species share some properties, such as having an abundance of omics data and literature information available to study them. *M. tuberculosis* [6] and *M. pneumoniae* [7] are considered model organisms. These organisms were not only selected within this thesis based on their societal relevance but were also selected as suitable for Systems Biology approaches due to the abundance of data we have on them. Analyses that integrate multiple omics data, as well as comparisons between model predictions, such as predictions on gene essentiality and experimental results such as transposon mutagenesis essentiality data, is of course only possible if such data is available.

Here we would like to briefly introduce the above-mentioned species. In **Chapters 2 and 3** we studied *M. tuberculosis* and identified three virulence strategies as well as the environmental and regulatory cascade that controls these three virulence strategies. *M. tuberculosis* is an obligate intracellular human pathogen [6] of great societal relevance. It is estimated that *M. tuberculosis* was responsible for 1.3 million deaths in 2021 in non-HIV infected patients and is one of the top 10 leading causes of death worldwide [8]. Although the relative occurrence of multi-drug resistant and

extreme drug resistant *M. tuberculosis* infections have not increased in the last 2 years [8], their treatment is expensive and lengthy with an average of 9 months [8]. *M. tuberculosis* is a special bacterium with a highly impervious waxy coating of mycolic acids and a very long replication time. This waxy coating as well as the ability of *M. tuberculosis* to enter a dormant near metabolic inactive state, make it resistant to many drugs [9], [10] and resilient against many stresses encountered in the human host. *M. tuberculosis* uses multiple strategies such as immune modulation, immune modulation, dormancy and phagosomal rupture [11]. Better understanding of how *M. tuberculosis* interact with its human host can help the development of new Tuberculosis vaccines or systems medicine approaches to shorten treatments, reduce costs and human suffering. One of the main problems in treating *M. tuberculosis* is its ability to adapt and enter a dormant state in which it is nearly immune to all kind of drugs. Understanding the environmental cues that regulate the three virulence strategies of *M. tuberculosis* and the switch to dormancy, could for example be used in treatments to trick *M. tuberculosis* to reactivate prematurely, leading to quicker bacterial clearance.

In **Chapter 4** we compare various *Staphylococcus* and *Streptococcus* species. These species include the well-known opportunistic pathogen *Staphylococcus aureus* which kills around 20.000 people in the US yearly [12], as well as *Streptococcus pneumoniae,* a commensal of the nose and oral cavity, responsible for killing millions a year worldwide as opportunistic pathogen [13]. *Staphylococcus* and *Streptococcus* species can grow in aerobic environments but are facultative anaerobes that produce lactate as part of their fermentation pathway. There are also various animal and some plant pathogens among *Staphylococcus* and *Streptococcus* species, such as *Streptococcus suis* [14] and *Streptococcus agalactiae* [15] which can infect humans. The ability of some pathogenic bacteria to jump from a host to humans is called zoonosis and is referred to as 'zoonotic potential' within this thesis.

In **Chapter 5** we analyse the central carbon metabolism of *M. pneumoniae* and in **Chapter 6** we predict the host and tissue specificity of various *Mycoplasma's*. *Mycoplasmas* are minimal intracellular and extracellular pathogens [16]. *Mycoplasmas* can stimulate their own uptake by host cells and primarily survive by evading the hosts immune response [17]. The *Mycoplasma* species group includes many commensals and opportunistic animal, plant and human pathogens [18] including intracellular pathogens such as *M. hyorhinis* [16] as well as pathogens such as *M. hyopneumoniae* which can be survive both intracellular and extracellular [19]. *M. pneumoniae* is an extremely slow growing organism with a doubling time of 20-60 hours [20]. *M. pneumonia* does not have a cell wall, instead it has an external cell membrane depended on the exogenous supply of fatty acids such as phosphatidylcholine, cardiolipin, phosphatidic acid and phosphatidylglycerol, sphingomyelin, glycolipids, cholesterol and diacyl-glycerol from its host [21].

1.3 Bacterial strategies to cause infection

Illnesses caused by bacteria are complex and involve many molecular interactions that are part of strategies deployed by both the host and by the bacterial invader. Like in 'The Art of War' by Sun Tzu [22], there are many strategies to beat one's opponent.

Some bacteria, like *Mycoplasma*, use a stealth approach by being as small, minimal, and slow growing as possible. *Mycoplasma* mimic their host on their surface in the hope to slip by the defences. Other bacteria take the offensive and try to overstimulate the immune response with toxins, such as some *Staphylococcus* and *Streptococci* bacteria, causing severe conditions such as toxic shock syndrome [19]. Yet, others like *Mycobacterium tuberculosis* can switch between several strategies such as mimicking the host on their surface, modulating the immune response, creating pores in the phagosomes of the macrophages cells that envelop them or going in a dormant like state for decades before re-emerging. Although bacterial pathogens use combinations of different strategies, there are similarities, overlaps in their strategies and overlap in the molecular building blocks that facilitate these strategies. As such, understanding one pathogen can help improve understanding of another pathogen. Pathogenic genes and the knowledge of their function is, in many cases, transferable between pathogens.

Of great societal relevance and interest to this author are the abilities of some bacterial pathogens to pass from animal or insect to human. These are called zoonotic pathogens. There are both bacterial and viral zoonotic diseases. Some examples of the zoonotic viruses are the HIV virus which originates from chimpanzees [23], the corona virus COVID-19 [24] which is suspected to originate from bats [25], [26] and more recently the 2022 outbreak of the Monkeypox virus [27] which originates from monkeys, chimpanzees and various rodents [28]. From these zoonotic viruses, Especially HIV and COVID-19 have caused great societal disruption and human suffering. Also many bacterial zoonotic pathogens have great impact human society such as *Streptococcus suis* [14] and *Streptococcus agalactiae* [15] which can be transferred from pigs and aquatic species to humans respectively. A more ancient and deadly zoonotic intracellular bacterial pathogen is *Yersinia pestis* [29], which caused bubonic plagues for at least 5000 years killing large parts of the European population during 14th–18th century [30], [31]. The impact of viral and bacterial zoonotic pathogens on society is great. Animal pathogens comprise and endless reservoir to the ever expanding pool of pathogens that can infect humans [32]. This ever-expanding pool of pathogens combined with increased multi-drug resistance and extreme drug resistance of both human and animal bacterial pathogens, is one of the great challenges humans must overcome in the 21st century [28]. The pool of known and unknown antibiotics is limited. The "WHO has declared antimicrobial resistance to be one of the top 10 global public health threats" in their 2022 report [19]. Understanding the complex interplay between human and bacterial pathogens, to better understand zoonosis, as well as to find systems approaches to combat pathogens in synergy with traditional antibiotics, is of great relevance nowadays.

To understand the complex interactions between pathogenic bacteria and their host, large amounts of *genomics*, *transcriptomics*, *proteomics*, *metabolomics*, and *phenomics* data are needed. These five '*omics*' data deal respectively with information on genes in the DNA, mRNA used to transcribe these genes, proteins that are synthesised based the mRNA, metabolites synthesized by enzymes and the phenotypic properties of bacteria. For the rest of this thesis, I will refer to these data types jointly as '*omics data*'. Since the whole is more than the sum of its parts systems approaches are needed to integrate and analyse all these omics data and to gain greater understanding of the biological system they represent. The field of Systems Biology

emerged to study combinations of these various omics data and the complex systems they represent since the whole is more than its parts [33].

1.4 Need for Systems Biology to understand infections

Holistic approaches that utilise multiple *omics* data are needed to understand complex systems such as the interactions between humans and bacterial pathogens. Systems Biology is a field that emerged naturally to deal with the increasing abundance of molecular and biological data as well as the increasing awareness of the complexity of life. The Human Genome project which started in 1990 and was finished in 2003 [34], can be seen as both a Bioinformatics or a Systems Biology projects. Since the project required large scale collaborative efforts, standardization, and computational approaches to deal with the vast abundance of sequencing data, the project paved the way for Systems Biology. Systems Biology is an important shift in paradigm, where opposed to reductionistic approaches that focusses on the parts, the system as a whole is studied.

Systems Biologist recognized that that the whole is greater than the sum of its parts. Properties such as emergence [33], robustness [35], modularity and oscillation [36] can emerge from network motives such as feedforward and feedback control systems. Systems Biology uses a holistic approach to identify such properties [33], [37]. For some the word 'holistic' is associated to 'vagueness' and 'pseudoscience'. Nothing could be more wrong though. Systems Biology is a highly transdisciplinary field using advanced qualitative and quantitative scientific methodologies, data integration, statistics and (mathematical-) modelling to gain greater understanding of biology [38]. Systems Biology is differently defined by different experts in the field. Some approach it from chemistry, physics or biology perspective, however, most including this author agree systems biology is at its core an engineering approach applied to the study of biological systems [39], [40]. Systems Biology requires tinkering and detective work as we demonstrated in **Chapters 2 and 3** and like in engineering, standardization of tools and data is key to the success of any project Systems Biology. A great example of such standardization is FAIR data management which aims to make data Findable, Accessible, Interoperable and Reusable [41]. We will further discuss the importance of proper standards for Systems Biology in the section FAIR and interoperable data. Systems biology comprises different disciplines and a rich set of tools and methodologies out of which the scientist must select the right combination to solve the problem at hand while considering the available data.

1.5 The systems biology toolbox

Similarity and guilt by association

Inference based on similarity is the most used tool in bioinformatics and systems biology. Most genes and proteins are assigned functions based on similarity to genes and proteins with a known function. All chapters in this this thesis use protein annotation by searching proteins with domain signatures in the PFAM protein family database [42]. Similarity can be based on DNA or protein sequence alignment. More commonly nowadays, more advanced methods such as similarity based on HMMs

models [43] of protein domains are used. Examples of guilt by association are the inference of a function or property based on connections in a graph, being part of the same cluster or close distance in a network. Similarity can be based on many aspects such as "underlying encoding, references to biological entities, quantitative behaviour, qualitative behaviour, mathematical equations and parameters and network structure" as defined by Liebermeister and Waltemath et al. [43] For example, when comparing models, one can use references to biological entities such as two reactions in two models involving metabolites with the exact same identifiers, means these reactions are likely to be similar. Underlying encoding, such as the SBML model version and level, means that models might be easier to compare than models from different modelling versions. Models with similar network structure, mathematical equation and similar parameters are likely to model the same pathways, possibly from the same organism.

Pattern recognition

Biology contains patterns that are associated to certain properties. Examples of patterns are a) positive and negative feedback loops in signalling and metabolism b) regulatory motifs in the DNA or RNA which are patterns that bind specific transcription factors, proteins that orchestrate which proteins are to be expresses in a certain condition c) patterns that capture a protein domain. Examples of tools that use pattern recognition in their basis are domain annotation software such as InterProScan [44] which annotates protein domains by searching for protein signatures in various protein signature databases such as the PFAM [42]. Protein domain models are built by identifying structural similarities in the multiple sequence alignments of proteins with high sequence similarity. Similarly, tools like MEME [43] detect binding motifs in the upstream regions of co-expressed genes based on multiple sequence alignment.

Network analysis

Using various *omics* data, networks of interactions can be mapped. Networks can represent interactions, connection, or similarity between different entities such as metabolites, proteins, genes, or messenger RNA. For example, a network could be based on protein-protein interactions or similarity in gene expression based on mRNA data, or correlation between metabolites concentration in various conditions. Networks provide an 'unordered' graphical representation of these system and can help identify clusters, modules, or patterns, in a graph. With unordered, I mean networks can be ordered using various network ordering algorithms but do not use a reference map or fixed reference layout. A cluster in a metabolite-metabolite correlation network can for example, be used to identify feed forward and feedback loops to be implemented in biochemical network models [40]. Clusters in transcriptomics data can be used to identify clusters of co-regulated genes [45]. Networks and heatmaps are often used to identify similarity based on clustering. Networks based on similarity score can however be calculated using many different similarity algorithms, each with their advantages and disadvantages. For example, Pearson Correlation and cosine similarity are invariant to scaling, which is very useful when working with metabolite data or text mining where you search for similarity in patterns, not similarity in absolute values. Additionally, Pearson correlation has the

added advantage of being invariant to adding a constant. This means similarity is still discovered if there is a systematic error in the measurements of some metabolites [46]. However, Pearson's Correlation evaluates the linear correlation between entities. In case of non-linear relationships Spearman rank-order correlation would be better at detecting similarity. More advanced methods such Context Likelihood [47] and Probabilistic Context Likelihood of Relatedness (PCLRC) which extends upon CLR, work even better when detecting non-linear relationships between metabolites [48]. Metabolite networks can appear linear, since many metabolites are linearly produced, and their concentration roughly linearly increase with small changes in substrate concentrations or enzyme concentration. However, metabolism in its essence is non-linear, since biochemical reactions are constrained by the Michaelis–Menten kinetics of their enzymes, resulting in hyperbolic enzyme reaction curves or sigmoidal reaction curves in reactions that involve allosteric control [49]. Hence, trying to capture these nonlinear relationships using Pearson correlation is sub-optimal. Explicit modelling using differential equations as demonstrated in chapter 6, is more fitting to capture non-linear correlations from metabolite data [50]. With the above examples, we hope to illustrate that there is no single method that works best for detecting similarity, since different methods have different strengths and weaknesses in defining similarity.

Different software for visualising networks is available. In this thesis we used a Synchronous Network Data Integration framework (SyNDI) [51], and its predecessor software for Data Integration Visualization and Analysis (DIVA) [52], to simultaneously visualise multiple omics network of *M. tuberculosis*. Visual exploration of simultaneously displayed omics networks is a very powerful tool for gaining knowledge of a biological system and for hypothesis building. In **Chapters 2 and 3** we use networks based on literature and multiple omics data to identify modules and regulatory binding sites associated to *M. tuberculosis* pathogenesis. In **Chapter 6** we use a graph based on correlation between abundance of different metabolites in metabolomics data to identify pathways that might be limiting for growth of *M. pneumoniae*. In **Chapters 4 and 5** we use co-occurrence of domains belonging to a biological functional group in other organisms to infer domain functional group annotation.

Modularity and mapping

Large complex systems can be broken down into functional units which are referred to as 'modules'. For example, a group or proteins that are regulated together and form an iron uptake system are together defined as a module called 'iron uptake system'. Similarly, complex metabolic pathways can be broken down into modules based on the thermodynamics and allosteric control that separate them from other modules. Modularity can be based on the combination of some properties, such as similarity in location, expression, or physical interaction between proteins. Ideally, one would like modules to be completely decoupled and independent from one another, especially if one would like to use a module in Synthetic Biology approaches [49]. One example of using modularity are maps of metabolism such as the Roche Applied Science 'Biochemical Pathway's map [53]. Visual maps, such as the Roche 'Biochemical 'pathway map, are important to understand the complex systems they depict. This is reflected in the increased integration of visualisation tools in modelling environments

used in Systems Biology such as PathwayTools [54] and Escher [49] and model platforms such as Seek [55] and Bigg model database [56]. Visualisations and maps create order in the chaos, and are in this authors opinion, undoubtedly one of the most helpful tools in the Systems Biology Toolbox. The saying, a picture is worth a thousand words, also holds true in Systems Biology. In **Chapter 2** we demonstrate how we build a modular map of *M. tuberculosis* pathogenesis where we identified modules based on gene co-expression in transcriptomics datasets, physical interactions among proteins, co-occurrence of gene or protein names in abstracts of scientific literature and functional similarity based on gene functional annotation. An example of where we identify modularity within this thesis is the detection of modules in central carbon metabolism by extensively analysing a dynamic model. Iterative rounds of model simulation and parameter estimation lead to the identification of modules in **Chapter 6**. Defining such modules in biological systems puts artificial boundaries and should not lead to the illusion of true decoupling and independence of these systems. In my opinion, balancing a reductionist approach that emphasis order, and a holistic approach that recognizes there are modules and emerging properties within a system, is what lead to the greatest understanding of biological systems. Creating modules and maps has been shown to be a useful simplification that help to increase understanding of otherwise unfathomably complex systems.

Mathematical modelling

Mathematical modelling such as constraint-based GEnome scale metabolic Models (GEM's) [55] or dynamic models of metabolism [55] are used to understand metabolism and to identify bottlenecks in metabolism. Dynamic models have the advantage of explicitly modelling enzymes kinetics, being able to capture and predict the dynamic behaviour on changes to internal or external stimuli, regulation as well as systems properties [57]. Living cells often contain thousands of reactions. Modelling such large systems dynamically is currently only possible by collapsing pathways and simplifying their representation to nearly linear reactions. Choosing modelling equations as well as fitting parameters to data is a non-trivial task [58]. Dynamic models are extremely data hungry, requiring large amounts of quantitative metabolite time-series data, even larger amounts of steady state data for parameter estimations or enzyme kinetic measurements. GEMs on the other hand require much less data and are great for constraint-based modelling, where one wants to optimize the yield of biomass, or a biomass associated product. GEMs are used to simulate fluxes through a metabolic network at a steady states using Flux Balance Analysis (FBA)[59] while dynamic FBA can be used to simulate a broader range of dynamic condition [60]. Although GEMs are oversimplified models, especially when a product is directly related to growth such as is the case of the biomass, they are very effective in predicting maximal possible yields and substrate utilization. An example of the use of GEMs in this thesis can be found in **Chapter 4** and an example of dynamic modelling of metabolism, using differential equations, can be found in **Chapter 6**.

"Machine learning and AI models", are often referred to as black box models that are used to predict properties by training and testing them on certain data [61]. The black box part means the models themselves do not mathematically resemble the structure of the systems they try to simulate or classify and that makes it harder to understand

why a black box model made a certain classification [61]. Despite being 'black box' models, machine learning models can be very useful for classification and can in many cases still be analysed to identify which features are important for the prediction of specific classes [61]. Furthermore, some machine learning models such as Classification Trees and Random Forest models are relative straight forward to interpret. An example of machine learning models to predict phenotypic models can be found in **Chapters 4 and 5**.

1.6 FAIR and interoperable data

As discussed in previous paragraphs, biological systems are complex and involve many types of data which scientist like to abbreviate as '*omics*' data. Great progress has been made to standardize data generation and storage in the life sciences. Examples of such progress are harmonization of data and model standardization as part of the EU COMBINE project [62], improved infrastructure for FAIR data managements such as the European Open Science Cloud (EOSC) [63], the incorporation of FAIR data management as part of the FAIR funding model [64]. Other examples of such progress are the use of semantic web technology for genome annotation pipelines [65], community amplicon analysis [62], computational modelling [66] and web-based cataloguing and sharing heterogeneous scientific research datasets, models or simulations, processes and research outcomes using FAIRDOM-Seek [67]. Furthermore, data generation is increasingly standardized, automated, and large scale. Large scale highly standardized data enables the use of AI in the discovery of biopharmaceuticals [68], personalised medicine [69] and the quantitative analysis of bacterial communities [70].

FAIR data management principles are defined as making data Findable, Accessible, Interoperable and Reusable [41]. FAIR data management enables scientist to reuse data, explore data and answer more complex questions more efficiently. FAIR data management is also very important for the social accountability of using public research funds. By making data FAIR by design, the chance of that data being re-used in future scientific studies increases dramatically. Findable and Accessible research data is rather common nowadays, with many peer-reviewed journals requiring access to the research data for publishing. However, the Interoperability and Reusability of data are often still lacking [71]. In order for data to be Interoperable and Reusable, data needs to be both human and machine interpretable [41]. This requires the data to be standardized in annotation, in data format as well as in the ontologies [72] that structure the data. FAIR data management often involves semantically stored data as Research Description Framework (RDF) [72]. RDF standardizes the storage of any type of data in the most rudimentary form*, subject-predicate-object* triples [66]. The properties and structure of RDF data is defined by one or more Ontologies. Semantic data has the benefit of being easy to query, even multiple databases are interoperable due to the use of standardized data format, identifiers, and ontologies [61]. This makes it possible to easily query over multiple databases. Semantic data enables scientist to freely explore multiple datasets, ask complex questions, or use machine approaches to find new correlations, with very little effort. This means that semantically stored FAIR data can benefit science and society for a much longer time than one time use

generated data. In this thesis I used a semantic annotation pipeline to annotate data, queried data from various external databases which use semantic data annotation in **Chapters 4 and 5**. I stored experimental data and modelling data from one of our projects in a semantic database as discussed in **Chapter 6** to adhere to the FAIR data management principles.

1.7 Thesis objective and outline

The overarching objective of this thesis is to improve understanding of interactions between bacterial pathogens and their various hosts. The main research question I try to answer is:

"What are the patterns in bacterial pathogen host interaction?"

Sub questions that I will address within the various chapters are:

1. *What are the strategies used by various pathogens to cause illness?*
2. *What are the strategies a model organism like M. tuberculosis deploys to infect the host?*
3. *How do functional groups of proteins associate to differences in pathogen host interaction?*
4. *Which genes confer zoonotic ability to bacteria?*
5. *Which genes determine the host and tissue specificity of bacterial pathogens?*
6. *What are the properties of M. pneumoniae central carbon metabolism to adapt to different environmental conditions?*

The various chapters use common concepts and a portfolio of strategies and methodologies to address the biology of pathogens and their interaction with their hosts. Below, I list the different chapters of this thesis.

In **Chapter 2** I create a visual and modular overview of the three virulence strategies of *Mycobacterium tuberculosis (Mtb)*. In this study I integrated literature information and available *omics* data to come to a system understanding of *Mtb*. I produced a visual map of the regulation of the three major virulence strategies as well as smaller maps of the complex regulatory systems of these three virulence strategies. These maps highlight the identified regulatory cascade that controls the different strategies in pathogen host interaction in response to environmental stimuli, such as the availability of divalent metals.

In **Chapter 3**, we provide examples of how Synchronous Network Data Integration framework (SyNDI) was used to identify two *espACD* associated stress clusters and their regulatory binding site and how we identified the sigma factor, SigE, as regulator of a sub cluster of the group of genes commonly regulated DevR, the dormancy regulated. In this work, regulated genes were identified using and iterative approach of motif identification through Meme [43] and motif scanning or matching by Fimo [73].

In **Chapter 4** I performed a multi-omics comparison of *Staphylococcus* and *Streptococcus* bacteria using genomic, transcriptomic data and transposon mutagenesis data, to uncovers pathogenic traits such as '*zoonotic potential*'. In this

chapter I used genomic and transcriptomic data as well as transposon mutagenesis essentiality data and various methodologies such as PCA, tSNE, phylogenetic trees, heatmaps, Genome Scale Metabolic modelling, and Random Forest classification. Jointly, these methods helped to identify interesting clusters of bacteria, their phenotypic traits as well as the proteins associated to these phenotypic traits.

In **Chapter 5** I integrated genomic and physiological data to predict which hosts and tissue various *Mycoplasma* bacteria infect. In this chapter I identified possible proteins responsible for the ability of some *Mycoplasma's* to opportunistically infect humans when having access to the blood stream, as well as various protein factors associated to host and tissue types.

In **Chapter 6** I present a dynamic model of central carbon metabolism of *Mycoplasma pneumoniae* to identify bottlenecks and metabolic dependencies. This dynamic model was used to investigated central carbon metabolism with the objective to improve growth of *Mycoplasma's* for optimal vaccine production. Robustness was identified as an inherent property of *Mycoplasma pneumoniae* metabolism and two main control hubs in central carbon metabolism were identified. Via analysis of metabolomics data, I identified some potential metabolic dependencies of *M. pneumoniae* on its human host.

In **Chapter 7** I look back and discuss the various successes, failures and bottlenecks encountered in this thesis. Additionally, I discuss the future perspective of the various methodologies used this thesis.

Chapter 2
Regulation of three virulence strategies of *Mycobacterium tuberculosis*: A success story

Adapted from:

N. A. Zondervan, Jesse C. J. van Dam, Peter J. Schaap, Vitor A.P. Martins dos Santos, Maria Suarez-Diez. "Regulation of Three Virulence Strategies of *Mycobacterium tuberculosis*: A Success Story". *In International Journal of Molecular Sciences* 19(2) 2018.

2.1 Abstract

Tuberculosis remains one of the deadliest diseases. Increased prevalence of multi and extensively drug resistant *M. tuberculosis* strains makes treating tuberculosis increasingly challenging. To develop novel intervention strategies, detailed understanding of the molecular mechanisms behind the success of this pathogen is required. Here, we review recent literature to provide a systems level overview of the molecular and cellular components involved in divalent metal homeostasis and their role in regulating the three main virulence strategies of *M. tuberculosis*: immune modulation, dormancy, and phagosome escape. We provide a visual and modular overview of these components and their regulation. Our analysis identified a single regulatory cascade for these three virulence strategies that respond to limited availability of divalent metals in the phagosome.

Keywords: *Mycobacteria, virulence, immune modulation, dormancy, escape, phagosome, divalent metal, pore, cAMP, manganese, iron, zinc, esx*

2.2 Introduction

Mycobacterium tuberculosis (*Mtb*) is the most successful known intracellular pathogen infecting roughly one third of the world population and killing about 1.3 million people in 2017 alone [74]. Treating *Mtb* infection is increasingly difficult due to increasing number of drug resistant, multi drug resistant, and extensively drug resistant strains [74]. To come up with new drug targets and treatment strategies, there is an urgent need to understand the molecular mechanisms supporting the success of this versatile pathogen. Here, we will review the regulation of three important survival strategies of *Mtb*: immune modulation, dormancy and phagosome escape [9], [75], [76].

Firstly, *Mtb* is a master in immune modulation. Its ability to interfere with host cell signalling-pathways allows it to carefully balance production of cytokines involved in activation of the pro-inflammatory and anti-inflammatory response [77], [78]. By balancing the pro- and anti-inflammatory immune response, *Mtb* delays phagosome maturation, harvests essential nutrients, and stimulates the formation of alveolar macrophage-dominated granulomas that shield it from more effective immune cells [79]. Secondly, when residing in the hypoxic granuloma, *Mtb* enters a metabolically near inactive and non-replicating dormant state in which it is immune to most types of drugs [80]. *Mtb* manipulates the macrophages to accumulate lipids, providing it with the nutrients required to sustain dormancy for multiple decades [79], [81]–[84]. Thirdly, *Mtb* has a highly regulated pore formation system that it uses to escape from the phagosome into the cytosol, resulting into necrosis of the host cell and dissemination of the bacilli [85], [86].

The fine-tuned regulation of these three virulence strategies is what makes *Mtb* such a successful pathogen. A large body of literature exist on these virulence strategies and the molecular components that constitute them. However, there have been few attempts to provide a systems wide overview of these three virulence strategies, their molecular components, and their regulation. Divalent metals play an important role in the regulation of some key aspects of these strategies [87]–[89]. Here, we will

present an overview of their involvement in this regulatory process. Detailed inspection of available knowledge pinpoints a single regulatory cascade as a main control hub for these three-virulence strategies, representing their interconnectivity as subsequent stages encountered in pathogen host interaction. A modular overview of the molecular components involved in divalent metal homeostasis and their components involved in the three virulence strategies can be found in Supplementary Files 1 and 2. In the following, we will discuss these components and the environmental cues that control them, and we will highlight the role of divalent metals in the phagosome.

2.3 Divalent metals at the interface of *Mtb* host interaction

Divalent metals such as iron, zinc, and manganese are required for proliferation and survival of all living organisms. Divalent metals appear, in all living beings, nearly exclusively as constituents of proteins and act as cofactors in many essential enzymes and environmental sensors [90]. Iron is the most commonly used divalent metal cofactor [90]. Iron containing enzymes are involved, among other processes, in electron transfer, maintaining redox balance and detoxification [91]. Manganese has the strongest affinity for ATP and is the preferred cofactor in cAMP production [92], [93]. Zinc plays a vital role as cofactor for numerous enzymes and DNA binding proteins, and serves as a structural scaffold for several proteins [94].

To prevent growth of bacteria, the host uses high affinity iron binding proteins such as lactoferrin, ferritin and transferrin, to keep the concentrations of free iron in the blood low, in the so-called iron sparing response [88], [95]. These proteins also bind other divalent metals such as manganese, albeit with lower specificity than iron. Similarly, calprotectin functions as high affinity calcium binding protein but also binds manganese, zinc and iron in the blood [96]. During infection, macrophages withdraw approximately 30% of the total circulating iron from the blood stream to restrict their availability, making macrophages environments rich in divalent metals [97]. Some intracellular pathogens use this defense mechanism to their advantage by stimulating phagocytosis by macrophages to get access to divalent metals and other nutrients. *Mtb* specifically targets alveolar macrophages which are rich in divalent metals while having reduced bactericidal abilities compared to other macrophages [84], [97].

Upon ingestion by a macrophage, *Mtb* is engulfed in a special compartment called the phagosome, in a process known as phagocytosis. The phagosome then fuses with vesicles containing enzymes and other proteins that facilitate the bacterial digestion process. Phagocytosis is a rapid process leading to phagosomal-endosomal fusion in approximately 3-4 minutes, acidification of the phagosome within 23-32 min and fusion with lysosome in 74-120 minutes, based on experiments with epithelial macrophages [98]. However, *Mtb* blocks phagosome maturation in an early phase leading to fusion with early endosomes and a pH of approximately 5.5 [99].

The macrophage continuously exports divalent metals out of the phagosome via Nramp1 and Nramp2 in a pH dependent manner. Many cell types express Nramp2 while only macrophages express Nramp1. Nrmap1 is mechanistically similar to Nramp2 but has a much higher specificity for manganese (Mn) compared to Nramp2

[89], [99], [100]. Mn is required as cofactor for the bacteria to break down oxidative compounds produced in the phagosome such as H_2O_2 [88], [92], [101]. Thus, restricting Mn availability in the phagosome by recruitment of Nrmap1 is an essential defence against intracellular pathogens. Nramp2 functions optimally around pH 6, a condition found in the early phagosome while Nramp1 has an optimal activity at a pH of 4.5 [89], [101]. Nramp1 is recruited to the membrane of maturing phagosomes and is associated with enhanced recruitment of the vacuolar V-H^+-ATPase -positive endosomes and/or lysosomes, resulting in acidification of the phagosome from pH 6.5 to 5.5 [99], [102]. Nramp2 is regulated separately from Nramp1 and co-localizes with transferrin receptors to early endosomes as well as with V-H^+-ATPase which provides the electro-genic force needed for Nramp1 and Nramp2 to operate [103], [104]. Thus, metal availability in the phagosome is tightly regulated by the host through the combined action of Nramp1 and Nramp2. Blocking phagosome maturation is an effective strategy to create an environment in which *Mtb* can outcompete divalent metal export from the phagosome. *Mtb* uses special high affinity siderophores (mycobactin) to gain access to divalent metals from both extracellular transferrin and the intracellular iron pool [97].

Within *Mtb* iron, zinc and manganese homeostasis are regulated by IdeR, Zur (previously known as FurB) and MntR respectively [91], [94], [105]. Ligation of Fe^{2+} to IdeR and Zn^{2+} to Zur stabilizes the formation of dimers that have strong affinity to binding sites involved in suppressing the genes in their respective regulons [87], [91], [106]. MntR in *Bacillus subtilis* contains two manganese binding sites as well as a dimerization site similar to IdeR and Zur [107]. There is a significant overlap between IdeR, Zur and MntR regulated genes, see **Figure 1**. An overview of regulation of molecular component by divalent metal regulators, IdeR, Zur and MntR can be found in Supplementary Files 1 and 2. These three regulators each suppress the main operon of genes coding for the ESX-3 secretion system and associated PE, PPE and Esx proteins homologues of ESAT-6 and CFP-10 (EsxA and EsxB) [105]. We will further discuss the ESX-3 transport system in a section below. In the following sections, we will discuss main characteristics of genes regulated by Fe, Zn and Mn respectively.

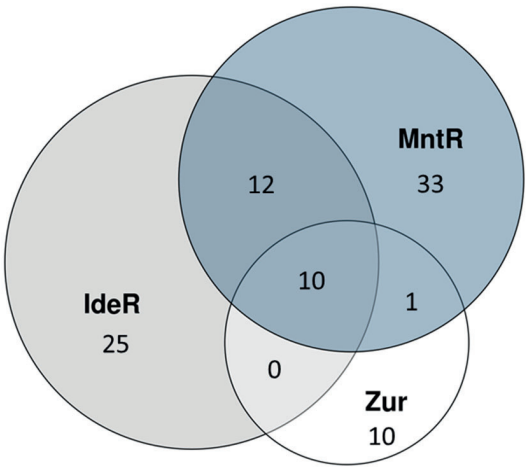

Figure 1. Number of genes in the IdeR, Zur and MntR regulons.

2.3.1 Iron homeostasis and redox sensing

Mtb produces high affinity hydrophilic and lipophilic siderophores, termed carboxy-mycobactin and mycobactin respectively. Mycobactin can bypass the phagosome membrane to scavenge iron from the extracellular iron storage protein transferrin [97], [108]–[110]. In addition, *Mtb* actively synthesizes deoxy-mycobactin during iron starvation [111].

Mtb combines the expression of a dedicated iron acquisition machinery with cellular components involved in immune modulation. By limiting acidification of the phagosome, *Mtb* maintains favourable conditions in which it can outperform active export of divalent metals by the macrophage's transporter Nramp1. *Mtb's* success in acquiring iron is illustrated by a 20 fold increase of iron concentrations in the phagosome between 1 and 24 hours of macrophage infection [104]. However, high iron concentrations renders *Mtb* much more vulnerable to the formation oxygen and nitrogen radicals upon phagosome maturation, as iron functions as a catalyst in the formation of radicals via the Fenton reaction [112]. Tight regulation of iron homeostasis is therefore essential, making IdeR an interesting drug target [113]. *Mtb* has adapted to deal with oxidative stress outside of the cell but is relatively vulnerable to endogenously generated oxidative stress in comparison to *M. smegmatis* [112]. Due to this vulnerability, vitamin-C is an effective drug to combat *Mtb* in the early stage of infection by inducing the Fenton reaction in iron rich phagosomes [114]. The oxidative conditions encountered in the phagosome leads to oxidation of the intracellular iron pool. Oxidation of the iron pool de-represses IdeR regulated genes among which some are involved in virulence. Upregulating expression of virulence genes in low iron and oxidative conditions is a common response in intracellular pathogens and has been observed in *Shigella dysenteriae, Corynebacterium diphtheniae, Yersinia pestis* and *Yersinia pseudotuberculosis,* as well as in *Mtb* [115], [116].

The iron pool within *Mtb* and the phagosome functions as redox sensor to the oxidative conditions encountered in the early phagosome. In oxidative conditions, ferrous iron (Fe^{2+}) is oxidized to ferric iron (Fe^{3+}) [117]. Ferric iron does not bind to IdeR, leading to upregulation of IdeR suppressed genes in oxidative conditions [113]. Genes suppressed by IdeR code for proteins involved in siderophore synthesis (*mbtA-G*), secretion (*mmpL4/5, mmpS4/5*) and uptake (*irtAB*) as well as 11 ESX-3 genes, among other [118]–[120]. Even though IdeR mainly functions as iron dependent repressor, IdeR also induces transcription of four genes. Among the induced genes, *bfrB* and, in a lesser extend *bfrA,* code for mycobacterial ferritin-like iron storage proteins which prevent overload of iron within *Mtb* [91], [121]. Analysis of the promoter region of *bfrB* revealed it contains two tandem IdeR binding sites involved in alleviating repression by Lsr2. Lsr2 is a histone like regulator that binds AT-rich regions virulence islands, including those coding for ESX-1, espACD and PDIM coding genes, acting as a global regulator to aid in the adaptation to extremes in oxygen availability [121]–[126]. Combined regulation of BfrB by Lsr2 and IdeR, suggests iron storage by BfrB is suppressed by Lsr2 during infection under changing oxygen conditions unless IdeR detects availability of intracellular ferrous iron which indicates a lack of oxidative

conditions. BfrA is required for efficient utilization of stored iron under low iron conditions, while BfrB is required for storage of iron under high iron conditions [127].

Iron homeostasis is an essential process for bacterial survival; therefore, its cellular components are interesting drug targets. This was shown in a knockout study of the *mmpS4/5* siderophore secretion, which resulted in limited intracellular availability of iron as well as intracellular accumulation of siderophores toxic to *Mtb* [128]. Another interesting drug target is HupB, a nucleoid-associated protein that protects *Mtb* against reactive oxygen species, regulates siderophore synthesis and was proposed to facilitate transfer of iron from ferri-carboxymycobactin to mycobactin [129], [130]. HupB stimulates transcription of its own operon in the absence of IdeR-Fe^{2+} [130].

IdeR also regulates genes involved in response to oxidative and acidic stress, among which the two-component system PhoPR. Two-component systems contain a histidine kinase sensor that senses specific environmental stimulus and a response regulator that gets phosphorylated by the sensor upon specific environmental stimuli. Many two-component regulators, among which PhoPR, also regulate their own operon [131]. Presence of multiple binding sites allows both positive and negative regulation depending on the concentration and phosphorylation state of the response regulator, as is the case for PhoPR [132], [133]. PhoPR is the main regulator of the oxidative and acidic stress response, but also is the initial step in a regulatory cascade controlling pore formation and phagosome escape. Six putative IdeR binding sites upstream of the *phoP-phoR* operon where located of which five where observed to bind IdeR in presence of iron [134]. Nevertheless, the exact role of IdeR in upstream binding of PhoPR remains to be determined.

Oxidation of the iron pool is also sensed by proteins containing iron-sulphur clusters such as the enzyme aconitase (Acn) and the regulators FurA and WhiB1-7. Acn catalyzes the isomerization of citrate to isocitrate via cis-aconitate in normal conditions. However, in low iron or oxidative conditions it binds to and suppresses translation of IdeR-mRNA while increasing translation of TrxC-mRNA [135]. The function of Acn as redox sensitive translational regulator is conserved in many organisms [117], [136].

FurA (ferric uptake regulator A) regulates the oxidative stress response by modulating expression of the operon coding for FurA and the KatG catalase [137]. KatG is essential for the breakdown of H_2O_2 radicals formed upon phagosome endosome fusion and activates the anti-cell-wall drug isoniazid. Recently, transcriptional activation of *furA-katG* was found to be regulated by RbpA, which is induced by H_2O_2 in a SigE dependent manner [138].

A third iron sensitive regulator is WhiB7. WhiB proteins are iron- sulphur cluster containing redox-sensing transcription factors. WhiB7 expression is auto-regulated by binding to its own promoter in response to antibiotics or redox stress [139]. An 80-fold upregulation of WhiB7 was observed upon treatment with antibiotics that bind to the 30S ribosomal subunit such as kanamycin and streptomycin [139]. WhiB7 is upregulated by iron starvation and was shown to induce transcription of *eis* and *tap* [140], two antibiotic resistance genes. Upregulation of *eis* increases secretion of IL-10

and slightly represses production of TNF-α by the host. IL-10 and TNF-α are involved in the anti-inflammatory and pro-inflammatory responses respectively [141].

In summary, oxidation of the iron pool is an important environmental cue to activate molecular components involved in iron sequestering, immune modulation, and virulence. IdeR, FurA, Acn, WhiB7, Lsr2 and SigE are all involved in the response to the oxidative conditions encountered in the phagosome and subsequent adaption through expression of a vast repertoire of molecules involved in iron homeostasis as well as genes involved in modulation of the immune response.

2.3.2 Manganese homeostasis and cAMP production

Manganese is one of the most abundant metal elements in nature [142]. Mn is involved in enzymes of diverse functionality such as photosynthesis as well as detoxification: Mn is used as cofactor for both synthesis and degradation of H_2O_2, superoxide and radicals [88]. The oxidative burst is a very effective bactericidal process to defend against intracellular pathogens such as *Mtb* and *Y. Pestis* [125], [143], [144]. As previously stated MntR is a regulator of Mn homoeostasis, however MntR is dispensable for *Mtb* growth in human and/or mice macrophages due to the limited availability of Mn in the phagosome. Manganese transport on the other hand is required for virulence and to break down oxygen radicals [105]. *Mtb* contains two superoxide dismutase's, SodA and SodC. SodA uses manganese as preferred cofactor and requires CtpC for metalation and export to the phagosome. Interestingly, *ctpC* transcription is induced in the presence of PhoP while *sodA* is predicted to contain upstream cAMP-CRP binding sites implicating it in its regulation [131], [145].

Another role of Mn we would like to discuss here is the Mn dependent activation of cAMP production in the early phagosome which was first proposed by S. Reddy *et al.* in 2001 [93]. S. Reddy and co-workers studied kinetics of membranes containing *Mtb* adenylyl cyclase CyA (Rv1625c). Their study revealed that the Michaelis-Menten constant (Km) for Mn-ATP is 70-fold lower than for Mg-ATP. This results in a 47-fold activation by 1mM Mn-ATP compared to 1mM of Mg-ATP at physiological conditions [93]. Mn is also essential for the CRP regulated, virulence associated type III phosphodiesterase Rv0805 [146], [147].

During infection, intracellular cAMP concentration increases ~50 fold and this is associated with a decrease in pH from 6.7 to 5.5 [148]. Among the 15 Adenylate Cyclases (AC) present in *Mtb* H37Rv, CyA has the highest measured cAMP production while AC (Rv1264) functions optimally at pH 6, which is the typically found at the early phagosome [148], [149]. *Mtb* was shown to secrete cAMP in a burst into the macrophage cytosol, resulting in a 10 fold increase in the host's TNF-α concentration, an important inducer of granuloma formation [150]. Rv0386 is needed for this cAMP burst [150].

The MntR regulon contains *mntH* (Rv0924c), coding for Mramp, an Nramp homolog that imports manganese (Mn) in a pH dependent manner; *mntABCD* (Rv1283c-Rv1280c) coding for an ATP dependent manganese transporter and *Rv2477c* coding for a manganese dependent ATPase which optimally functions at pH 5.2 [151]. Interestingly, Rv2477c was postulated to be involved in resistance to tetracyclines and

macrolides [151]. Additionally, MntR as well as Zur regulate *Rv2059-Rv2060* coding for two components of an incomplete ABC transporter of unknown function. *Rv2059*-Therefore, it is more likely that this transporter is involved in transporting other divalent cations like Co^{2+}, Cu^{2+} or Ca^{2+} to substitute Mn and Zn in some conditions. A second possibility is that this operon codes for a divalent cation exporter, to counter the side effect of unwanted uptake of divalent cations such as Cu^{2+} by the high expression of manganese and zinc transporters [105]. Manganese uptake plays an important role in virulence of many bacteria. For instance, supplementing *Salmonella typhimurium* with manganese prior to infecting macrophages, decreased its lethal dose 50 fold [152]. Similarly, manganese acquisition in the gut was shown to allow *S. typhimurium* and *Salmonella enterica* to evade neutrophil killing by calprotectin and reactive oxygen species, while patients with mutations in manganese transporter Nramp1 were shown to be much more susceptible to pathogens such as *Mtb* [92], [99], [125], [143], [153], [154].

MntR regulates WhiB6 which regulates *espACD* and some DevR (previously known as DosR) regulated genes [155]. DevR is the main regulator of dormancy and *espACD* is involved in pore formation [156] and will be discussed below. The WhiB6 Fe-S cluster is necessary for the negative control of the devR regulon and positive control of the ESX-1 secretion system, whereas apo-WhiB6 induces the dosR regulon and suppresses ESX-1 expression in *M. marinum* [156]. A model was proposed where holo-WhiB6 positively regulate ESX-1 operon while upon reaction with reactive oxygen species and NO, apo-WhiB6 and WhiB6-DNIC are formed respectively. Both apo-WhiB6 and WhiB6-DNIC activate DevR regulated genes to shift metabolism and maintain energy and redox homeostasis [156].

MntR interacts with toxin-antitoxin system RelJ and RelK in which MntR functions as antitoxin [157], [158]. Additionally, VapBC26, and VapB30 toxin-antitoxin system both requires Mg or Mn for their ribonuclease activity which to inhibits growth [159], [160]. These results indicate Mn might function as environmental cue in the regulation of growth.

2.3.3 Zinc homeostasis

The third and final divalent cation we would like to discuss is zinc, the only redox stable divalent metal of the three. As previously stated, zinc homeostasis is regulated by Zur (FurB), a Zn^{2+} dependent repressor. Zur knockout studies identified 32 genes that are upregulated in the *zur* knockout mutant of which 24 belong to eight transcriptional units that were shown to be directly regulated by Zur [94]. Zur expression levels are regulated by SmtB encoded for by *smtB,* an upstream gene which is co-operonic with *zur*. SmtB functions as a repressor, which is deactivated upon binding to Zn^{2+} [94].

There are three possible zinc uptake systems regulated by Zur. Firstly, Zur regulates the *sitABC* like genes (*Rv2059-2060*) which are also regulated by MntR that were previously discussed. This suggest that this transporter might function as Zn importer [92], [161], [162]. Secondly, Zur regulates *Rv0106* coding for a protein similar to the

B. subtilis putative zinc low-affinity transporter YciCas [161]. Thirdly, EsxG-EsxH proteins were shown to be able to bind zinc which might implicated them in zinc transport [163].

Other interesting targets of Zur are five genes coding for ribosomal proteins that can function in the absence of zinc, in contrast to their zinc dependent counterparts which normally bind to the 30S ribosomal subunits [94], [164]. Although Zur was found to be able to positively regulate some genes in other pathogenic bacteria via repression of non-coding small RNAs, no such regulation was found in a *zur* knockout *Mtb* mutant [87].

2.4 Three main virulence strategies of *Mtb*

The three virulence strategies discussed in this review, namely immune modulation, dormancy, and phagosome escape, represent subsequent stages in *Mtb*-host interaction. These strategies extend and complement each other, which is reflected in their regulation. While many pathogens directly express components involved in phagosome escape, *Mtb* keeps a low profile and activates key virulence strategies such as phagosome escape only when immune modulation fails, and the phagosome becomes inhospitable. However, immune modulation also complements phagosome escape and dormancy since immune modulation leads to conditions such as granuloma formation and cholesterol accumulation which is needed to prepare *Mtb* for dormancy and phagosome escape.

2.4.1 Immune modulation

Mtb uses a number of virulence proteins, complex lipids and secreted metabolites, to modulate the immune response and arrest phagosome maturation to prevent fusion with late endosomes and lysosomes [75], [148], [165]–[169]. In case of successful immune modulation, phagosome maturation is halted resulting in a pH of approximately 5.5 [99], [102]. The macrophage controls intracellular trafficking, including phagosome maturation, through 42 distinct Rab GTPases. Rab5 is associated with phagosomes immediately after phagocytosis and normally diffuses quickly, allowing Rab7 to associate to the phagosome, which allows fusion of the phagosome with lysosomes. Studies with *M. bovis* have shown that mycobacteria halts phagosome maturation, by blocking vesicle fusion between stages controlled by Rab5 and Rab7, with no Rab7 being accumulated in macrophages even after 7 days [167]. Similarly for *Mtb,* Rab7 was shown to be recruited to the phagosome but its premature release prevents fusion of the phagosome with late endosomes [165], [170].

In addition to the earlier discussed ESX-3 secreted proteins, several other proteins and molecules are involved in blocking phagosome maturation. Secreted tyrosine phosphatase (PtpA) is involved in the exclusion of the vacuolar V-ATPase preventing acidification and fusion with lysosomes [168], [171]. cAMP secreted by *Mtb* blocks phagosome lysosome fusion by inhibiting actin assembly [169]. Additionally, a number of virulence lipids interfere with the phagosome Golgi trafficking needed for maturation of the phagosome [170], [172]. Among these are trehalose monomycolate and dimycolate, phthiocerol dimycocerosate (PDIM), sulpholipid-1, diacyl trehalose,

and pentacyl trehalose. Of these lipids, PDIM was shown to play a role in phagosome escape and will be discussed in the section below.

Mtb is very successful in balancing the expression of molecular systems involved in activating the pro- and anti- inflammatory responses of the host to direct the immune response to favourable conditions for its survival. *Mtb* achieves this balance through multitude sensors and that integrate many environmental cues. One important family of regulators involved in sensing internal conditions are the iron-sulphur cluster containing WhiB family of regulators, already mentioned in the section on iron homeostasis. Different WhiB regulators have different redox potential and sensitivity to oxidative agents such as O_2 NO and for some, thioredoxin like protein disulphide reductase activity has been reported [139], [173]–[175]. Many WhiB genes are regulated by cAMP-CRP [139], as summarized in **Figure 2**.

WhiB1 is an essential regulator that senses NO, is regulated by cAMP-CRP and is associated with resuscitation [175], [176]. WhiB4 is associated to the oxidative stress response while WhiB5 is required for resuscitation [177], [178]. DNA binding has only been experimentally proven for WhiB1, WhiB2, WhiB3, WhiB6 and WhiB7 [139], [156]. Interestingly, WhiB1-3 are induced, upon nutrient limitation, by exogenous cAMP and during infection indicating they are involved in sensing the redox state of *Mtb* [179]. For WhiB1-3 it was shown that their DNA binding ability is enabled by NO by bringing their iron-sulphur cluster in their nitrosylated or apo-form [139], [180]. WhiB2 and WhiB3 are down regulated in presence of O_2 while others like WhiB3, WhiB6, and WhiB7 are upregulated in early or late hypoxic response. Of the WhiB genes, WhiB7 is most upregulated in the macrophage with a 13 fold induction while being 80 fold induced by antibiotics that bind the 30S ribosomal unit [174]. WhiB3 senses NO and O_2 via its iron-sulphur cluster [144] and regulates genes involved in assimilation of propionate, a byproduct of cholesterol degradation, into virulence lipids [181]–[184]. Virulence lipids regulated by WhiB3 include sulfolipids, diacyltrehaloses, and polyacyltrehaloses which result in both higher pro- and anti-inflammatory cytokine levels, and function as redox sync [45], [182]. WhiB3, PhoP and Lsr2 bind to and regulate the *whiB3* operon. MprAB might induce *whiB3* through upregulation of Rv0081 which was predicted to induce the *whiB3* operon [45]. In addition, WhiB3 together with DevSTR regulates expression of *tgs1* which is needed for the production of triacylglycerol, a storage lipid which without *Mtb* cannot resuscitate from dormancy [81], [144], [185]. WhiB1 is associated with resuscitation as it induces transcription of *whib1, rpfA, ahpC and Rv3616c groEL2* in the absence of NO upon upregulation of WhiB1 by cAMP-CRP [175]. Interestingly, WhiB1 also interacts with GlgB, which is essential for optimal growth of *Mtb*, by reducing intramolecular disulfide bonds [139], [175], [178].

For a full review of WhiB proteins we refer to the excellent paper by Larsson et al [174]. For a review of the function of WhiB like proteins and a network view of WhiB1-3 regulated genes and their connection to other virulence factors such as cAMP and CRP we refer to the review by Fei Zheng et al [139]. An overview of WhiB regulators and the environmental cues they respond to can be found in **Figure 2**.

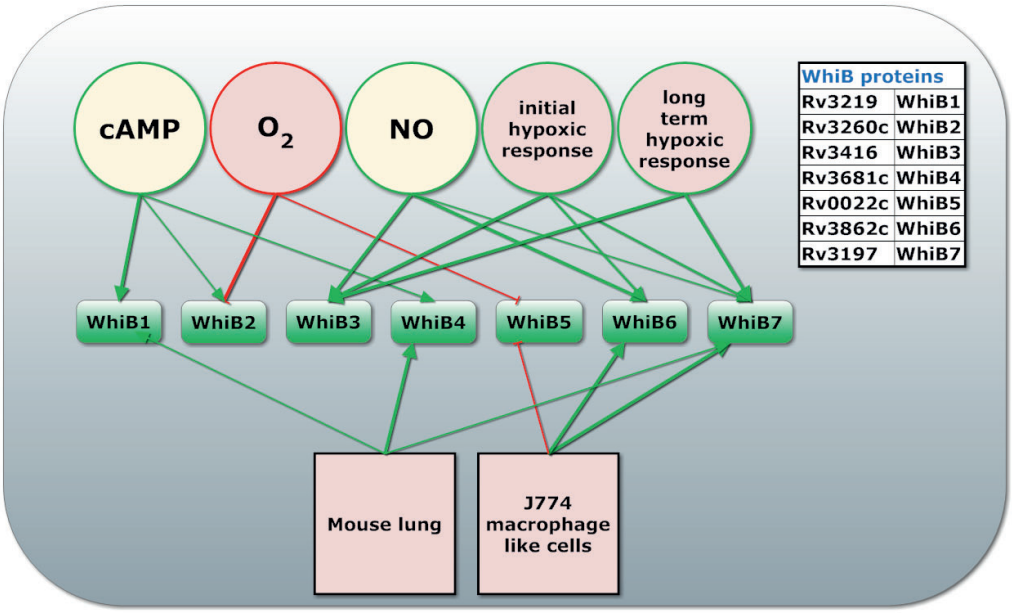

Figure 2. WhiB1-7 transcriptional response to environmental stresses. Proteins from the WhiB family are presented in the squares. The circles in the top indicate environmental cues (O2, NO, cAMP availability) or infection stages (initial or long-term hypoxic response). Squares represent different environments (mouse lung and JJ774 macrophage like cells). Arrows indicated regulation (green for induction, red for inhibition of transcription) with the line width indicating the strength of the interaction [139], [174].

Two highly regulated virulence systems are EspACD, involved in phagosome escape, and GroEL2, an abundant chaperonin involved in blocking apoptosis. Regulation of GroEL2 is summarized in **Figure 3**. GroEL2 is a highly antigenic gene associated with increased release of IL-10 and TNF-α which is also associated with cAMP secretion into the cytoplasm of the macrophage [148], [150], [169], [180], [186]. GroEL2 forms a dimer and is normally associated to the cell wall. However, Hip1 cleaves cell wall associated GroEL2 to form monomers that are able to cross the phagosome membrane and inhibit apoptosis by interacting with mitochondrial mortalin [187], [188]. In this way Hip1 modulates the macrophage responses by limiting macrophage activation and dampening the activation of TLR2-dependent pro-inflammatory responses [188]. Interestingly, Hip1 has also been reported to function as lipase, making the proteolytic function of Hip1 somewhat disputed [189]. *Mtb* inhibits apoptosis of the macrophage through aggregation of mitochondria around the phagosome and increased activation of mitochondria resulting in limited cytochrome C release, an important inducer of apoptosis [190].

CMR and HrcA positively regulate *groEL2* expression upon acidic and anaerobic stress [180], [191]. CRP induces *whiB1* expression in presence of cAMP while WhiB1 represses its own operon as well as *GroEL2* in the presence of NO [180], [192]. GroEL2 is therefore only expressed in the presence of CMR or heat stress or while NO is absent (See **Figure 3**). GroEL2 expression is induced 24 hours post infection, but not at 2 hours after infection while other CMR regulated genes, like *Rv1265* and *PE_PGRS6,* are induced at 2 hours post-infection [193].

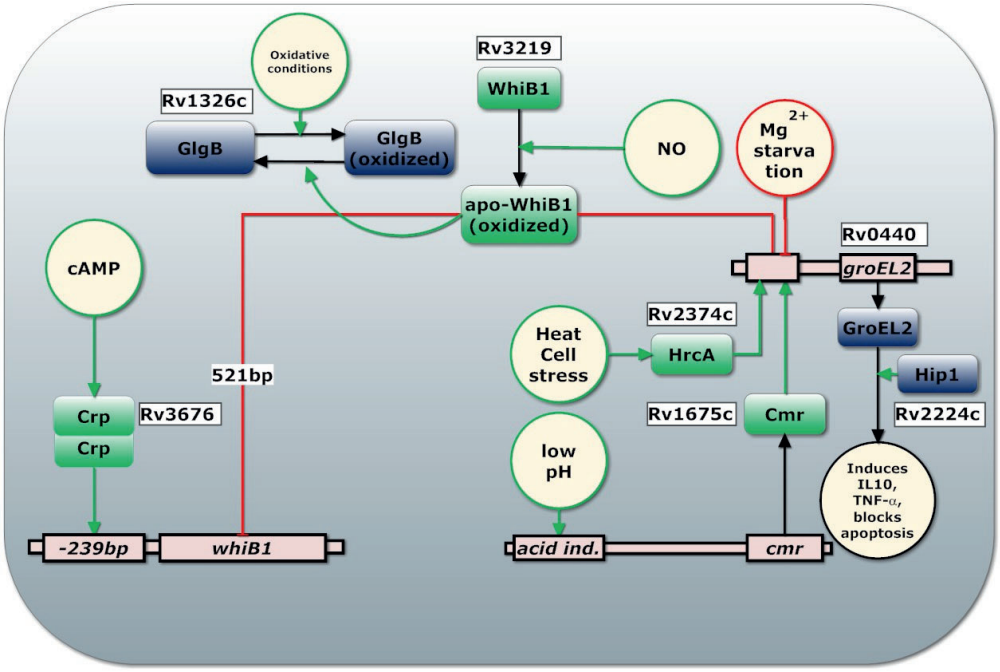

Figure 3. Regulation of GroEL2. Squares represent proteins, circles represent pools of simple chemicals, environmental cues or factors. Green lines indicate induction of transcription while red lines indicate inhibition of transcription. Black lines indicate causal effects.

2.4.2 ESX-3 secretion system

The ESX-3 secretion system is the only one of the five ESX systems that is essential for *in vitro* growth of *Mtb* [194], [195]. ESX-3 is involved in divalent metal homeostasis and immune modulation. ESX secretion systems are specialized secretion systems for the transport of extracellular proteins across the hydrophobic, and highly impermeable, cell wall of *Mtb* [166], [196].

Regulatory binding site for all three divalent metal regulators IdeR, Zur and MntR can be found in the ESX-3 core operon promoter [119], [163], as summarized in **Table 1**. The triple control of ESX-3 might allow *Mtb* to switch partly to other divalent metals in the absence of one of these three. This hypothesis is supported by the observation that siderophore knockout mutants low in iron contain much higher zinc concentrations [104]. However, many ESX-3 associated genes are regulated by only one or two of these regulators, indicating dedicated roles in homeostasis of specific metals [197].

Table 1. Suppression of ESX-3 core genes and associated genes by IdeR, Zur and MntR.

Gene	***IdeR***	***Zur***	***MntR***
esx3-operon[1]	1	1	1

esxG-esxH	1	1	1
esxQ	0	1	0
esxR-esxS	0	1	1
esxW	0	0	1
ppe3	0	1	1
ppe4-pe5	1	1	1
ppe9	1[2]	0	0
pe13	0	0[3]	1
ppe19	0	0	1
ppe20	0	0	1
ppe37	1	0	0
ppe38	0	0[3]	0
ppe48	0	1	0
pe_pgrs61	0	0	1

1Rv0282-Rv291
2Positively regulated by IdeR
3Reported as Zur regulated by Maciag et al. [94]*Predicted not to be in the Zur regulon in another study* [52]

All three divalent metal regulators regulate EsxG and EsxH which play an essential role in secretion of PE and PPE proteins [197]. PE and PPE proteins have immune modulating properties and comprise nearly 10% of the coding potential of the *Mtb* genome [198]. A large number of studies exist on the immune modulating properties of ESX-3 secreted PE and PPE proteins [195], [197]–[203]. The ESX-3 secreted protein pair EsxG-EsxH, targets the endosomal sorting complex to impair fusion of the phagosome with the lysosomes, while increasing association with the endocytic pathway leading to fusion with transferrin containing vesicles [163], [166], [195]. PE5-PPE4 were found to be critical for the siderophore-mediated iron-acquisition functions of ESX-3 [197]. PPE38 Inhibits Macrophage MHC Class I expression, dampens CD8+ T-Cell responses and was shown to be required for virulence of *M. marinum* [202], [203]. PPE37 was found to reduce the production pro-inflammatory factors tumor necrosis factor alpha and IL-6 [200]. Pe_pgrs61 binds to TLR2 in a Ca^{2+} dependent manner, leading to increased IL-10 production, while PE5 and PE15 trigger activation of the host MAP kinases required for IL-10 production [198], [201]. IL-10 is an important anti-inflammatory cytokine. IL-10 reduces the expression of *iNOS,* limiting production of nitric oxide (NO) in the phagosome [195], [198]. Enhanced IL-10 expression plays an important role in inhibiting early protective immunity and blocking phagosome activation [204], [205]. In addition, a direct role for IL-10 in *Mtb* reactivation has been observed [204]. Interestingly, IL-10 also modulates lipid metabolism by enhancing uptake and efflux of cholesterol in macrophages [204]–[206]. *Mtb* is known to induce foamy macrophage using immune modulating proteins as well as secreted lipids to deregulate the macrophages lipid metabolism via the macrophage lipid-sensing nuclear receptors PPARγ and TR4 [84], [205]. One study reported observing *Mtb* to exploited host vesicle trafficking and lipid storage by the recruitment of iron bound mycobactin to lipid droplets which move to and discharge their content in the phagosome [108]. Another study found that *Mtb* uses membrane vesicles containing immune modulating molecules as well as Mycobactin to interact with the macrophage during infection [207]. Further research is needed to investigate

the proposed synergy between modulation of host vesicle trafficking, lipid acquisition and iron acquisition.

2.4.3 Phagosome escape and pore formation

The second main virulence strategy deployed by *Mtb* is phagosome escape. A model of regulation of pore formation can be found in Figure 4. ESX-1 and ESX-1 secreted proteins EsxA (ESAT-6) and EsxB (CFP-10) have been implicated in phagosome escape of many *Mycobateria* such as *M. marinum*, *M. kansii* and *Mtb* [208]–[211]. The virulence lipid phthiocerol dimycocerosates (PDIM) and EsxA from *Mtb* were shown to interact with the host cell membrane and in concert, induce phagosome membrane damage and rupture in infected macrophages [211], [212]. A recent study reported that many claims about pore formation at neutral pH are due to contamination with detergent from the washing step [76]. The same study found membrane-lysing capabilities for EsxA only to occur below pH 5, to be contact dependent, and accompanied by gross membrane disruptions rather than discrete pores. For the sake of simplicity, we refer here to the process of cytosolic access as *pore formation* although more research is needed to find out if cytosolic access is only achieved through lesions or also through formation of pores.

The ESX-1 secretion system is involved in secretion of virulence proteins among which those shown to be involved in pore formation and phagosome escape EsxA (ESAT-6) and EsxB (CFP-10), secretion associated proteins EspA-D, EspF and secreted immune modulating PE and PPE proteins [196], [213]–[215]. Although EsxB is the main pore forming protein, other ESX-1 secreted genes are required for EsxB secretion and proper functioning of the ESX-1 secretion machinery. EspD stabilizes the extracellular levels of EspA and EspC, and it is required for EsxA secretion but does not require ESX-1 for its own secretion [216]. Secretion of EspA, EspC, EsxA is codependent on each other, suggesting they might be secreted as a multimeric complex or that they are part of the secretion machinery itself [217], [218]. This theory is supported by a study showing that EspA forms dimers by disulphide bond formation after secretion; disruption of this disulphide bond affects cell wall stability as well as the functioning of the whole ESX-1 secretion system [219]. Recently, an EspC-multimeric complex was observed to form filamentous structure that could represent a secretion needle [220]. Inactivation of MyCP1 protease causes hyper-activation of ESX-1 while protease inhibition leads to attenuated virulence during chronic infection [221], [222]. A balanced activation and deactivation of ESX-1 through MycP1 proteolysis of EspB is required during chronic infection. MyCP1 and MyCP5 are required for stability of the ESX-1 and ESX-5 secretion complex respectively [223]. Without ESX-1, *Mtb* is unable to disrupt the phagosome membrane and make contact with the cytosol, leading to highly diminished pathogenicity [213].

ESX-1 and secreted factors EsxA and EsxB are regulated by the two-component systems PhoPR, previously mentioned. The importance of PhoP for virulence was confirmed in knockout studies that showed *phoP* knockout mutants to be attenuated in mouse bone marrow derived macrophages, lungs, livers and spleen [224]. A single point mutation in *phoP* in *Mtb* H37Ra decreases the DNA affinity of PhoP and strongly contributes to the reduced virulence of this strain [225]. PhoPR regulated genes are

upregulated in acidic and oxidative conditions. Recently studies show that PhoP interacts with SigE, which is upregulated in acidic pH and upon cell stress [226]. Additionally, polyphosphate was implicated to be needed for normal transcription of *phoP* as well as well as for transcriptional regulation of *sigE* by *MprAB*, although these results could not be reproduced [227], [228]. PhoP/R influences transcription of some 80 (according to some sources up to 150 [229]) genes directly as well as the transcription of a large number of genes indirectly via upregulation of WhiB6, EspR, DevS/R and WhiB3 [45], [131].

EspR is a transcriptional regulator upregulated by PhoP. EspR induces transcription of the *espACD* (*Rv3612-16c*) operon which is essential for escape from the phago(-lyso)some [216], [219], [230]. PhoP therefore controls, directly (*espB/E-L*) or indirectly (*espA/C/D*), the 13 Esp proteins secreted by ESX-1 [230]–[232] Recently it was found that holo-WhiB6 increases transcription of its own operon, the ESX-1 regulon and suppressed the DevR regulon, while apo-WhiB6 formed in anaerobic conditions and by prolonged exposure to NO, suppresses the ESX-1 regulon and induces the DevR dormancy regulon [156]. Interestingly, gene expression of EsxB by WhiB6 was highly induced after 30-min of NO exposure, decreased at 60 minutes and is highly reduced after 3 hours of exposure to NO, indicating a short but intense activation of *espACD* by holo-WhiB6. Additionally binding sites for WhiB6 and Rv0081, a transcriptional factor regulated by MprAB, were predicted upstream of *espACD* [155]. These results suggest WhiB6, which is induced by PhoPR and MntR, plays an essential role in the regulation of phagosome escape and dormancy.

Induction of transcription of *espACD* by EspR requires the presence of PhoP [230]. In addition, MprAB, Lsr2 and CRP bind to the promotor region of *espACD* operon. Lsr2 represses transcription of both the *espACD* and the ESX-1 operon [155], while CRP binding inhibits expression of *espACD*[233]. Lsr2 binds to AT rich regions in the DNA, mostly virulence genes and is required for adaptation to extreme oxygen conditions [124], [125]. We hypothesize it is likely that Lsr2 represses the operon containing ESX-1 genes and *espACD* in oxidative conditions. This could serve to avoid further aggravation of the immune response. MprAB functions as a repressor of the *espACD* operon in cellular stress conditions, however MprA/B is also required for full expression of *espACD*. It is plausible to assume both positive and negative regulation by MprAB occurs based on the presence of multiple binding sites for MprA and two transcriptional start in the *espACD* operon [155].

Like the post-translational activation of GroEL2 by HiP1, membrane lysing capability of EsxA is activated only upon dissociation of EsxA from EsxB in acidic environment (pH 4-5) encountered when the phagosome matures. Acetylation improves dissociation of EsxA from EsxB at higher pH, a model where acetylation leads to reduced virulence was proposed [234]. Interestingly, acetylation of proteins in *Mtb* is cAMP dependent [210]. Taken together, these studies indicate pore formation is strictly regulated, most likely only occurs when cAMP is depleted (no cAMP-CRP), might be inhibited by sudden changes in oxidative conditions (Lsr2), the phagosome acidifies and become hypoxic (PhoPR) and pore formation is transiently inducted by WhiB6 upon sensing NO[156]. MprAB further modifies activation of *espACD*, most

likely both positively upon initial cell damage and negatively after prolonged cell stress and accumulation of polyphosphate, as indicated in **Figure 4**.

It should be mentioned that in addition to their role as regulators, Lsr2, CRP and EspR have also been characterized as nucleoid-associated proteins and as such might serve additional functions such as structuring the organization of the chromosome and, as has been shown for ESX-1 and *espACD* operon to protect DNA region from oxygen radicals [124], [233], [235].

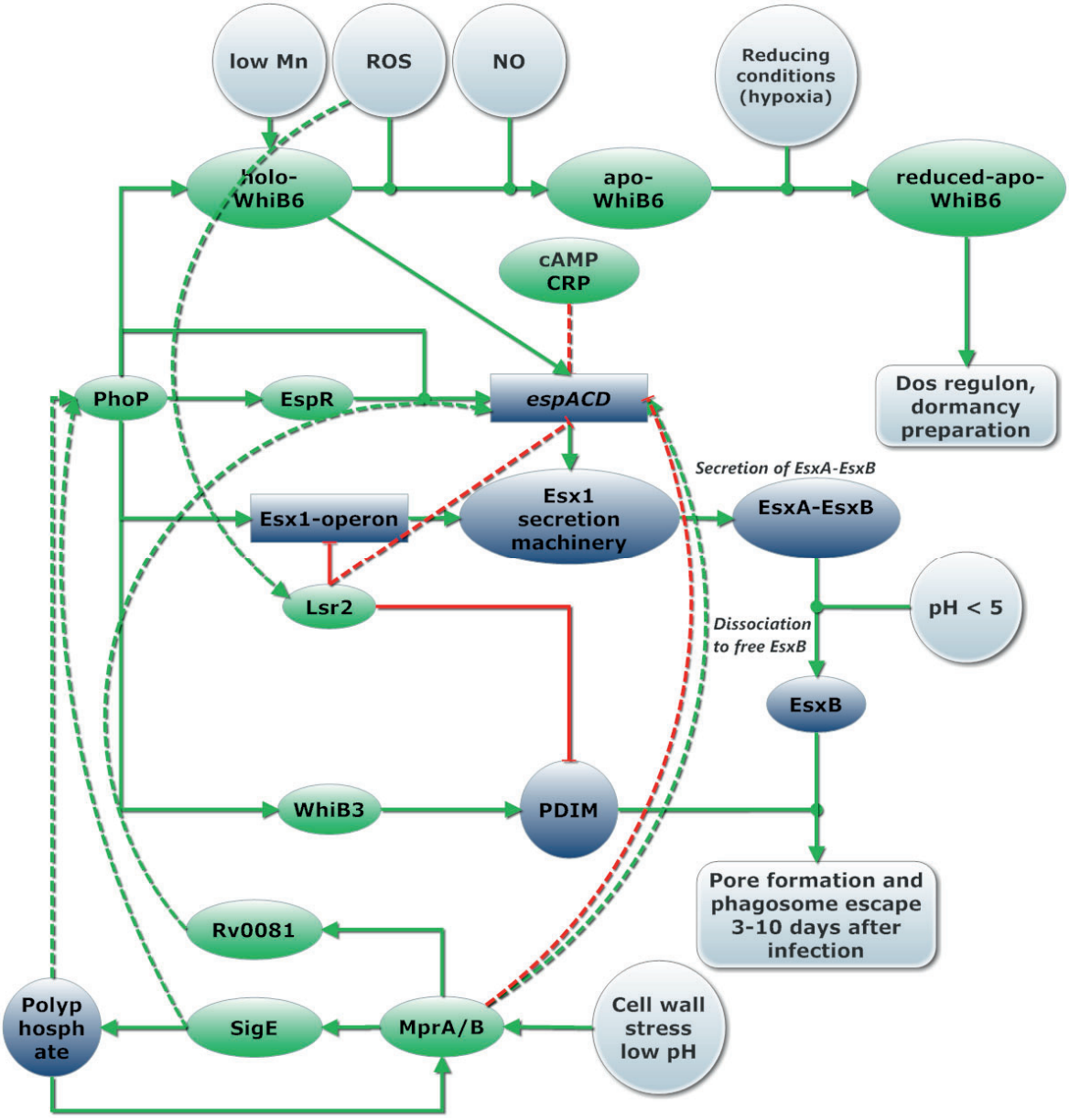

Figure 4. Regulation of pore formation. The circles represent environmental conditions. Arrows indicated regulation (green for induction, red for inhibition of transcription) with dashed lines for uncertain effects. Regulators are depicted in green, proteins and other molecules dark blue while operons are depicted in squares.

2.4.4 Dormancy and modulation of granuloma formation

The third virulence strategy deployed by *Mtb* is onset of dormancy. Dormancy is a non-replicating and metabolically near inactive state at which *Mtb* is immune to most drugs and can survive for decades [9], [81]. Dormancy occurs upon formation of mostly hypoxic granulomas [236]. Immune modulation that stimulates granuloma formation will therefore be discussed as a part of the dormancy virulence strategy.

When *Mtb* runs out of cAMP to secrete thereby suppressing phagosome lysosome fusion, the macrophages phagosome will fuse with late endosomes and lysosomes. As a result, the phagosome becomes increasingly hostile with lower pH, production of oxygen radicals and NO and fusion with vesicles containing lysozymes. In contrast, conditions encountered in granulomas are slightly more favourable for *Mtb*. Granulomas have reduced capacity to form oxidative radicals [83].

Mtb stimulates TNF-α production which leads to granuloma formation among others through secretion of cAMP into the cytosol [141], [204], [237]. A number of studies indicate that granuloma may be dispensable for preventing bacterial dissemination and may actually contribute to *Mtb* persistence and shield *Mtb* from more successful immune cells [79], [82], [83]. According to some models, *Mtb* containing granuloma's contain two types of macrophages: classically activated and alternatively activated[79]. *Mtb* shifts the macrophage population within the granuloma from being classically activated to alternatively activated macrophage which produce more anti-inflammatory cytokines (IL-10, TGF-β) and arginase, which compete with iNOS for the use of arginine as a substrate reducing NO production [79], [83], [238]. A balance of pro-inflammatory and anti-inflammatory response via stimulation of TNF-α and IFN-γ production is needed for granuloma formation while IL-10 is the main negative regulator for this response, inhibiting formation of dense and hypoxic mature fibrotic granuloma's [79], [204]. Moreover, parameter sensitivity analysis for a granuloma model, showed IL-10 had the strongest influence on myofibroblast numbers at 300 day post infection and indicated IL-10 to play a major role in preventing differentiation of immune cells needed to develop protective immunity [79], [204].

Several regulators allow *Mtb* to sense and adapt to hypoxia and maturation of the phagosome. The most important of these regulators is the two-component regulator DevRST which regulate genes coding for proteins that help *Mtb* prepare for dormancy and subsequent resuscitation [239]–[241]. A visual representation of DevRST response to environmental cues is present as part of Supplementary File 1. Both DevS and DevT can activate the DevR regulon through phosphorylation of DevR which autoregulates its own operon through cooperative binding to two binding sites [10], [240]–[242]. DevT provides initial activation of the DevR regulon through phosphorylation of DevR and has the strongest sensitivity to CO and a weaker binding to NO and O_2 compared to DevS. DevS is sufficient for DevR activation after 5 days of infection [243], [244]. DevS phosphorylates DevR even in the presence of small concentrations of NO, negatively regulates the DevR regulon through phosphatase activity in the presence of O_2 while positively regulating the DevR regulon in reducing conditions [243], [245], [246].

Interestingly, even under non-inducing conditions, and as such no phosphorylation of DevR, the DevR regulon is activated upon high enough concentrations of DevR, providing a possible explanation for enduring induction of the DevR regulon which might occur after prolonged autoactivation of its own regulon [10]. Among DevR regulated genes there are a few types of regulation. While some genes are strongly upregulated within a few hours of infection others are only mildly induced after 12-24 hours in hypoxic and high NO conditions [242]. DevR and other two-component regulators can fine tune expression of genes through the presence of multiple binding sites and through phosphorylation which stimulates cooperative binding [241].

CO is released by the enzymatic activity of heme oxygenase-1 (HO-1) in lungs infected by *Mtb* [247], [248]. CO is an important dormancy inducer. Interestingly, *Mtb* has a unique heme scavenging and degrading systems that does not produce CO allowing *Mtb* to degrade heme without inducing the immune response or its own dormancy regulon.

Interestingly, there is evidence for two DevR regulated proteins to be involved in stabilizing the 30S ribosomal units under hypoxic conditions, while slowing down translation and protein synthesis in the process [236], [249]. *Mtb* uses lipids such as cholesterol as primary nutrient in this phase of infection via genes regulated by KstR and IdeR [45][183], while upregulating production of TAG via *tgs1* which is under control of DevR and Whib3 [144].

Protein-protein interaction was observed between DevT and NarL, a lone two-component response regulator involved in nitrate and nitrite respiration in *Escherechia coli* [250]–[252]. Although the genes regulated by NarL in *Mtb* are unknown, we argue it is plausible that NarL is involved in regulation of *nirB, narU, narX, narU, nuoB* that are currently thought to be part of the DevR regulon.

NO is produced in the maturing phagosome and is an important dormancy cue sensed by DevT and DevS. *Mtb* expresses two truncated heme proteins, GlbN and GlbO, that help it detoxify from nitrate containing oxygen radicals such as NO while residing in the macrophage [253]–[256].

Interestingly, *GlbN* is co-*transcribed* with *lpRl* coding for Lipoprotein LprI, which Acts as a lysozyme inhibitor [257]. The *GlbN-lpR1* Activated isoniazid inhibits truncated hemoglobin N that protects against reactive nitrogen and oxygen species as well as AcpM, which is required for mycolic-acid production [87], [258]–[260]. NO was found to help *Mtb* to survive in hypoxic and acidic conditions through anaerobic respiration [252], [261]. In addition, nitrate respiration plays an important role in dormancy and protection against hypoxic and acidic stress [261], [262].

Although DevRST and WhiB3 are involved in the preparation for dormancy, the enduring hypoxic response measured in a *devR* knockout mutant showed 230 genes to be differentially expressed with roughly half of them upregulated in in the first day of hypoxia and the other half only upregulated at 4 and 7 days of hypoxia [263]. These results indicate many genes involved in the enduring hypoxia response are not regulated by DevR. Resuscitation from dormancy is more elusive and less studied than dormancy. Resuscitation involves ClgR and both SigH and SigE are upregulated upon

reaeration [264]. Also cAMP-CRP plays a role in resuscitation as it upregulates *rpfA* one of the five resuscitation promoting factors [192], [265], [266].

2.5 Success through tight regulation of virulence strategies

Mtb anticipates changes in the interaction with the host by upregulating both internal and external sensors and regulators involved in sensing progression of the immune response. This allows the bacteria to adjust more quickly to progression of the immune response. External sensors involved in survival in the macrophage consists mostly of two-component regulators [229] (such as DevRST, PhoPR, MprAB, SenX3-RegX3, NarL) while for internal sensors, WhiB family proteins and regulators such as CRP and CMR are used. These sensors and regulators appear interconnected, thus forming a single regulatory cascade that controls the three virulence strategies, as represented in **Figure 5**. This regulatory cascade integrates many internal (as cAMP, Mn, Mg, oxidative conditions, and presence of NO) and external environmental cues (phagosome pH or cell wall damage) for fine-tuned regulation of key virulence systems. Examples of such virulence systems downstream this cascade are GroEL2, ESX-1, EsxAB and EspACD. Pore formation by EsxA depends on the regulation of ESX-1 by PhoP, Lsr2 and WhiB6, and on regulation of EspACD by Lsr2, EspR, PhoPR, MprAB, WhiB6 and Rv0081. Post translationally, pore formation by EsxA is regulated by proteolytic activity of MycP1, acetylation of EsxA and dissociation of EsxA-EsxB upon acidification of the phagosome [85], [124], [125], [155], [156], [208], [210], [233]–[235]. Similarly, GroEL2 is regulated by CRP, WhiB1, HrCA and Mg^{2+} starvation and post-translationally regulated by proteolytic cleavage by Hip1 [180], [187], [188], [191]–[193].

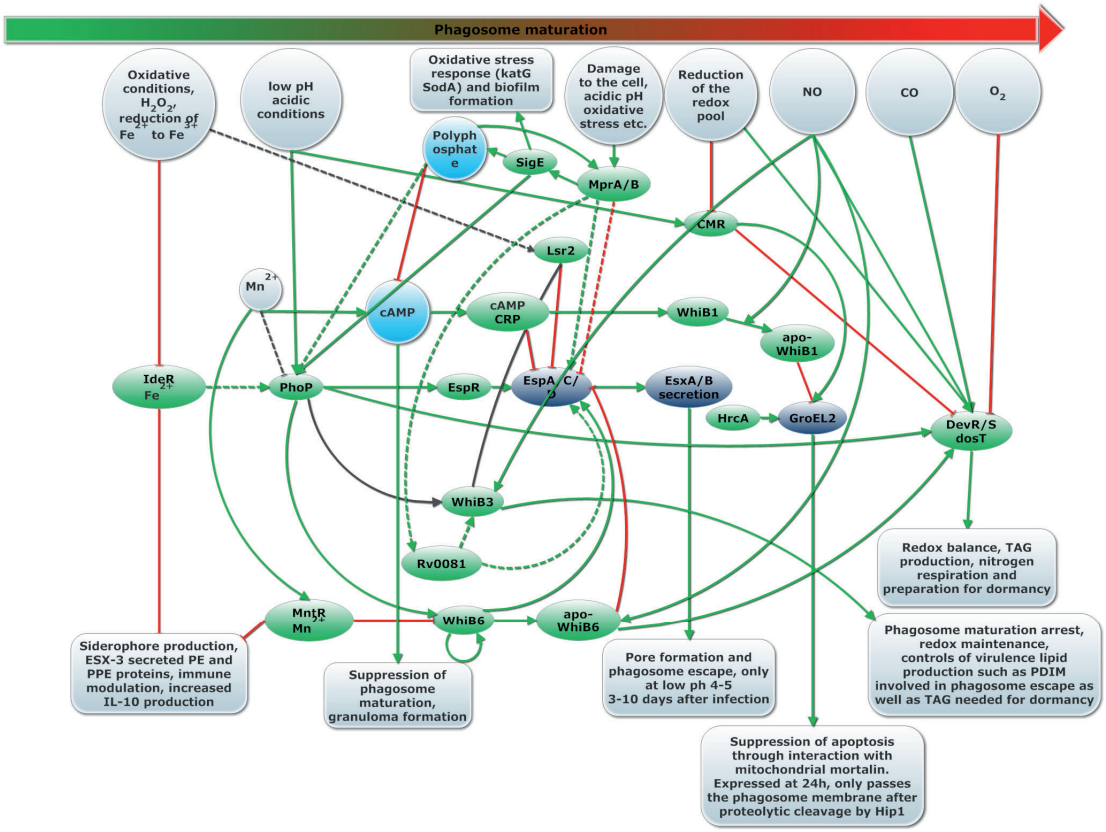

Figure 5. Overview of the regulatory cascade that integrates environmental cues to active the immune modulation, dormancy, and phagosome escape virulence strategies. Arrows indicated regulation (green for induction, red for inhibition of transcription) with dashed lines for uncertain effects. Regulators are depicted in green, proteins and other molecules dark blue while operons are depicted in squares. The large arrow on the top represents the progression of the immune response.

There is a great amount of overlap in this cascade, so that multiple environmental signals are considered in the regulation of these genes, as indicated in **Figure 5**. For example, some PhoPR regulated genes are predicted to have cAMP-CRP binding sites [267]. These genes are upregulated upon oxidative stress and low pH but suppressed in the presence of cAMP-CRP, as is the case for *espACD*[268]. Some PhoPR regulated genes are also regulated by DevRST, WhiB3 and by MprAB. An even larger overlap exists in genes regulated by DevRST and MprAB, indicating integration of CO, NO, hypoxia and cell stress in the regulation of these genes [269]–[271]. We argue that based on the overlapping regulation of the three virulence strategies, these strategies extend and overlap each other. The order of activation of these strategies is likely to vary depending on the dynamics between *Mtb* and the host. Timing of specific virulence strategies also vary for different *Mtb* strains [272]. Some strains gain cytosolic access within hours of phagocytosis while others require 3-10 days [85], [272].

Pore or lesion formation is linked to immune modulation. Cytosolic access is need for secretion of cAMP and other immune modulating factors, such as GroEL2, into the macrophage cytosol [272]. There are still many unanswered questions regarding the exact role and regulation of GroEL2. Firstly, it is unknown at which conditions

proteolysis of GroEL2 by Hip1 (*Rv2224c*) occurs. Secondly, Hip1 was reported to mainly function as lipase in one study [189], further research is needed to confirm whether GroEL2 is a direct substrate of Hip1. Strict regulation of GroEL2 suggests it to have an important role in virulence.

Interestingly, there are many parallels in regulation of virulence systems between *Mtb* and other pathogens. Understanding *Mtb* as one of the most successful intracellular pathogens can therefore provide insight in common strategies deployed by intracellular pathogens. For instance, positive regulation of virulence genes by PhoPR and suppression by cAMP-CRP appears to occur in more pathogens. In *Y. pestis,* PhoP directly binds to and transcriptionally activates *crp* and *cyA* leading to merging of the PhoPQ and CRP-cAMP regulon [273]. Similarly, a major virulence island is positively regulated by PhoP while being suppressed by cAMP-CRP in *S. typhimurium* [274]. In *Mtb,* PhoPR regulates pro-inflammatory virulence genes such as the ESX-1 operon as well as genes involved in protecting against oxidative stress, when cAMP is depleted. cAMP does not only suppress phagosome maturation but also acts as an internal sensor of phagosome maturation, through pH dependent secretion of cAMP.

Some aspects in the regulation of PhoPR and cAMP in *Mtb* require more research. Firstly, the function of multiple IdeR binding sites upstream of the *phoPR* suggests complex regulation of the *phoPR* operon by IdeR and thus by iron bioavailability. Secondly, the exact cue for activation of PhoP remains unknown. Upregulation of *phoPR* in acidic conditions has been observed as well as under Mg^{2+} starvation, however this later observation could not be reproduced [181]. Transcriptional analysis of *Mtb* showed many genes in the PhoPR regulon to be upregulated during the first hours of infection (20 min to 2 hours) while the phagosome acidified from pH of 6.5 to pH 5.5 [137]. PhoPR stimulates expression of *aprABC*, an important regulator of the intracellular pH [181]. These results indicated PhoPR directly or indirectly senses pH. Recently, it was discovered that PhoP interacts with acid inducible extracytoplasmic sigma factor SigE, providing a possible explanation for activation of the PhoP regulon at low pH [226]. Extracytoplasmic sigma factors provide a means of regulating gene expression in response to various extracellular changes, hence their name.

Secondly, we argue entrance of *Mtb* in the early phagosome is likely to lead to higher abundance of Mn. Pathogenic *Mycobacteria* species such as *Mtb* and *M. avium,* have high manganese concentrations at 1 and at 24 hour after infection compared to non-pathogenic *M. smegmatis* [104]. Mn availability might also be affected by Mramp, a pH dependent Mn H^+ symporter with maximal activity between pH 5.5 and 6.5 matching the conditions found in the early phagosome. Mn is an important cofactor for cAMP synthesis, and it is likely to increase cAMP production in the early phagosome. cAMP-CRP and PhoPR co-regulate virulence genes directly or via regulators such as WhiB6, which is linked to Mn deficiency. Based on the strong affinity of PhoP for Mn we hypothesize Mn might play a role in both cAMP and PhoPR regulation [92], [154]. Depletion of Mn and secretion of cAMP might lead to de-repression of cAMP-CRP suppressed genes such as *espACD* as well as activation of these genes through PhoPR.

Thirdly, polyphosphate is needed for optimal PhoP activation [227]. Polyphosphates are potent inhibitors of type III adenylyl cyclases in *M. bovis* which agrees with the opposing roles of cAMP-CRP and PhoPR in respectively inducing genes involved in the anti- and pro-inflammatory response in *Mtb* and other pathogens. Polyphosphate is implicated in the activation of PhoP and is part of one of two positive feedback loops in the regulation of *mprAB* and *sigE* [226]–[228]. Polyphosphates kinase production is conserved in all bacteria and is associated to induction of dormancy and activation of virulence genes in many pathogens [275]. Knockout polyphosphate kinases *ppk1* mutants, have reduced biofilm formation, are more susceptible to drugs and are impaired in growth in guinea pigs [227], [276]. Interestingly, SigE is involved in regulation of polyphosphate. MprAB and SigX3-RegX3, induce transcription of *sigE* upon cell wall stress or phosphate starvation, while anti sigma factor RseA binds to and neutralizes SigE in reducing conditions [277], [278]. RseA is degraded by ClpC1P2-dependent proteolytic activity depending on its phosphorylation by the eukaryotic-like Ser/Thr protein kinase PknB [278]. SigE, polyphosphate and MprAB are involved in a double positive feedback loops through polyphosphate and ClpC1P2 of which a visual model is provided by Manganelli *et al* [278]. Polyphosphate functions as phosphate donor for MprAB under low ATP condition. Additionally, SigE regulates the transcription of the *furA-katG* operon in response to oxidative stress in *Mycobacteria* [138]. SigE knockout strains are strongly attenuated and a recent study shows a *sigE* knockout strain provide an even more effective live vaccine than BCG [279]. Taken together, these studies indicate SigE plays an important role in adapting to low pH, cell wall and oxidative stress through upregulation *furA-katG,* activation of some PhoPR induced genes, MprAB and inhibition of cAMP-CRP through polyphosphate production. The interplay of SigE, polyphosphate and the hypothesized role of Mn in PhoPR and cAMP regulation should be further investigated.

Another aspect we want to address is the link between IdeR, cAMP, cholesterol degradation and phagosome escape. IdeR, KstR KstR2 co-regulate the cholesterol degradation pathway in *M. bovis* [183]. We suggest a similar synergy between IdeR regulation and cholesterol degradation in *Mtb*. Transcription of cholesterol degradation genes in *Mtb* is dependent on the presence of CyA [280]. Regulation of cholesterol degradation by IdeR and cAMP would suggest access to cholesterol is associated to the initial stage of *Mtb* host interaction when the iron pool is oxidized and cAMP is produced to avoid phagosome maturation. Interestingly, EsxA and other pore forming toxins specifically inserts themselves into phosphor lipid (phosphatidylcholine) and cholesterol-containing liposomes [234], [281]. Giant foamy macrophages rich in cholesterol are at the center of *Mtb* containing granuloma's that turn necrotic [79], [83], [84], [205], [281]. Accumulation of cholesterol was shown to be essential for uptake of *Mtb* by the macrophage [282]. Additionally, cholesterol was shown to increase association of TACO, a coat protein that prevents degradation of *Mycobacteria* upon fusion with lysosomes [282]. We argue that accumulation of cholesterol in macrophages not only increases *Mtb* survival in the phagosome by serving as carbon source, but also might assists in its escape from the phagosome.

In summary, in this review we provide an overview for understanding divalent metal homeostasis and their role in regulating three essential virulence strategies of *Mtb*: immune modulation, dormancy and escape. Sensors of environmental and internal

cues, including divalent metal availability, form a single regulatory cascade that controls these three virulence strategies. The role of polyphosphate, cAMP and manganese in this cascade requires further investigation.

2.6 Supplementary Materials

Supplementary Materials: The following are available online at http://www.mdpi.com/1422-0067/19/2/347/s1. All Supplementary files, Figures as well as additional code not present in the supplementary files of the published manuscript are available at: https://github.com/NielsZondervan/PhD Thesis.

2.7 Acknowledgments

This work has been supported by European Union through the SysteMTb project (HEALTH-F4-2010-241587), the Horizon 2020 research and innovation programme under grant agreement No. 634942 (MycoSynVac) and the FP7 programme under grant agreement No. 305340 (INFECT).

Chapter 3
Deploying a Synchronous Network Data Integration framework do identify gene regulatory motifs in *Mycobacterium tuberculosis*

Adapted from:

Erno, Lindfors*, Jesse C. J. van Dam*, Carolyn Ming Chi Lam, **Niels A. Zondervan**, Vitor A. P. Martins dos Santos, Maria Suarez-Diez. “SyNDI: synchronous network data integration framework”. In *BMC Bioinformatics* 19(1) 2018.

*Equal contributions

3.1 Introduction

Mycobacterium tuberculosis (Mtb) is responsible for an approximate 1.3 million deaths in non-HIV infect patients in 2021 and is one of the top 10 leading causes of death worldwide [8]. *Mtb* evolved from a group of ancient pathogens [283], adapting its persistence virulence systems over thousands of years to become one of the most successful bacterial pathogens [284], [285]. Its long evolutionary relationships with humans made *Mtb* adjust to an obligate intracellular pathogenic lifestyle using various pathogenic strategies such as immune modulation [286], dormancy [9] and phagosomal escape [272]. Although *M. tuberculosis* is considered highly resistant to horizontal gene transfer, there is evidence that during its evolution large scale gene deletion and horizontal gene transfer took place [287]. For example, some virulence proteins such as EsxA and EsxB can be found in other pathogens such as *S. aureus* [288] and *Streptococcus suis* [289]. Other systems such as the poly-polyketide metabolism of *Mtb* and *Yersiniae pestis,* show some similarities and are vulnerabilite to the same drugs [51] while 19 *Mtb* genes coding for polyketide synthesis are suspected to be of Eukaryotic origin [290]. The diversity in strategies and molecular building blocks to cause virulence makes *Mtb* a good model organism of bacterial pathogenesis. The vast amount of omics data available for *Mtb* make *Mtb* suitable for Systems Biology Approaches [6], [52], [291]. Systems Biology looks at the system at large, for example, by building large networks based on genomics, transcriptomics, proteomics, or metabolomics data. Biomolecules such as genes, proteins and metabolites are represented as nodes while their interactions are represented as edges. Edges represent different types of interactions or associations depending on the studied biomolecules and the considered datatypes. For example, a network of transcriptional similarity would show genes with similar expression patterns over a range of conditions. A network based on protein-protein interaction would show physical interaction between proteins, while in some cases the network shows more general associations such in the case of STRING db [51] that includes physical interactions and predicted functional associations. Additional information, such as the presence of common motifs in the upstream regions of the considered genes or experimentally validated associations can further strengthen the reliability of the obtained networks.

Multiple methods have been developed to infer networks from different types of omics data [292], [293]. and integrative approaches have been developed to generate consensus networks combining the strengths from the multiple methods through a wisdom of the crowds approach [51]. In addition to combining all information in a consensus network, an alternative approach is to synchronously browse multiple networks including the consensus network, as sometimes similarities and associations can only be seen in one or a few omics data types by applying dedicated methods and approaches. Synchronous Network Data Integration framework SyNDI [51] and its predecessor DIVA [52] provide a framework to allow simultaneous visualisation, selection and browsing over multiple networks. Networks can be either generate top-down from experimental data using various algorithms or constructed bottom-up from biological pathway databases such as the Wikipathway database [294]. These multiple network data integration frameworks include an approach called

Meme2Fimo [52]. Meme2Fimo combines iterative motif elicitation in the upstream regions of genes found in close neighbourhood in a reference network, using MEME [43] and the identification of additional genes in the same network region also harbouring the elicited motif, using FIMO [73]. The iterative motive elicitation and matching is complemented with network neighbourhood considerations thereby leading to more accurate predictions on regulatory mechanisms.

ESX-1 is a type VII secretion system required for the secretion of virulence proteins such as EsxA (ESAT-6) and EsxB (CFP-10). These are involved in immune modulation and phagosome escape [11], [85], [295]. EspACD is required for EsxA-EsxB secretion and pore formation [52], [216]. Multiple regulators such as PhoP, EspR, MprA, CRP are involved in modulation of ESX-1 and its secreted factors [233]. The transcription factor DevR mediates the hypoxic response of *M. tuberculosis* and triggers the onset of dormancy which enables long term survival of the bacteria within the lung granulomas of the human host [296]. DevR regulon is essential for persistence and pathogenesis of *M. tuberculosis* [297]. ChipSeq experiments initially identified over 600 gene targets for DevR [45]. Integration of heterogeneous molecular networks with this data led to the identification of five groups of genes with distinct expression profiles among this initial set [52]. Here, we present the exploration of the regulation of ESX-1 associated genes espA, C and D and the role of DevR in regulating these genes and the identification of additional biding motifs associated to DevR in *M. tuberculosis* using the SyNDI framework. This analysis has previously been briefly presented to illustrate the use of SyNDI (Lindfords et al.) and here we present a more extended version.

3.2 Materials and Method

3.2.1 Meme2Fimo

Meme2Fimo [52] is a tool to iteratively identify a motif and search for genes that contain a motif in their upstream binding site in combination with network mining. We can summarize this method as a three-step protocol. For more technical details we refer to the method section within the SyNDI paper [52].

1) The approach starts by manually selecting an initial cluster of genes that are in the neighbourhood of each other in a network of interest. For example, a cluster of genes in a co-expression network. The selected cluster of genes is used in identifying a motif using Meme [298].
2) The second step is to use Fimo [73] to locate any other occurrences of motifs identified in step one within the complete genome. All occurrence of similar motifs in the upstream binding region of genes are scored and ordered according to their p-value. Genes with a low p-value indicate the found motif found upstream is more like the motif that was searched for. Similar motifs, with a p-value below a cut-off value, can be used as input for a new round of motif building. Step one and two can be repeated until the motif search returns no new genes below the p-value cut-off, meaning that subsequent iterations would not change results or until the researcher is satisfied with the identified motif and its associated genes.

3) The third step is to identify sub-clusters within a cluster of genes. There are two ways to identify such sub clusters. First, one can remove the expression data from conditions that are associated to a motif found in a cluster of genes, and generate a new co-expression network without this data [52]. The second way, which used within this chapter, is to deselect genes that are already associated to a motif.

Both MEME [43] and Fimo [73] are used with default settings. Upstream regions considered contain intergenic regions of up to 1000b. In some cases, a naïve operon extension is applied where all consecutive genes in the same strand are in an operon, if the intergenic distance is less than 1000b.

We combined this approach with literature search as well as searching in online motif databases manually to see if the motifs are associated to a known regulator. For a visual representation of the Meme2Fimo workflow, see **Figure 6**.

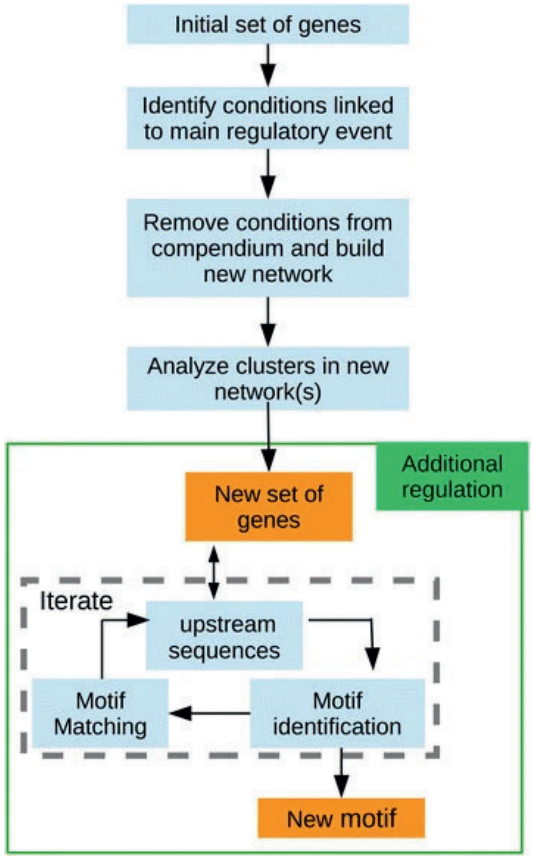

Figure 6. Pipeline to uncover additional regulatory layers. Step 1: Identify conditions linked to the main regulatory event for the initial gene set. This can be done using bi-clustering techniques or by direct comparison with the expression levels of the regulator (if known). Step 2: Build co-expression networks in the remaining conditions. Alternatively, instead of building a new network, the researcher can deselect genes with a known upstream binding motif. Step 3: Identify the closest neighbours of the selected genes in the new networks. Step 4: iterative round of motif identification/matching to identify the secondary motif and the set of genes with this motif in their upstream regions [43].

3.2.2 Multiple networks

A pipeline for the reconstruction of gene co-expression networks from a compendium of expression data was described in [52] to where we refer the reader for additional details. This pipeline is highly customizable, and its default values correspond to the following brief description. From a gene expression compendium, similarity between gene expression profiles is scored using Pearson's correlation for each gene pair. The significance of the similarity is scored using an estimate for the null model based on the rest of the similarity scores obtained for the members of the pair evaluated independently [47]. A generalization of the data processing inequality is iteratively applied to prune possible spurious associations from the network [299]. Stand-alone scripts implementing this pipeline can be retrieved from Additional file 3 of [52].

Here, we used SyNDI framework to synchronously represent and explore the following networks built on *M. tuberculosis* data.

In our explorations we use the following network views.

- CLR [47]: Co-expression Network using the Context likelihood of relatedness algorithm
- STRING.db fusion: STRING network based on fusion between proteins in other bacteria
- STRING.db: STRING Neighbourhood network, indicates genes are frequently occurring in the same genomic neighbourhood, which indicates genes might be involved in the same function or pathway.
- (bbh) networks: operon and BLAST based homology network, shows duplicated genomic regions and homology.

For a description of the pipeline that generated these networks, we refer to the article by J. van Dam et al. [52]. In the following sections we will describe for each exploration path details on the genes that were selected for each round of iterative motif searches. Annotation for genes in the results section were retrieved from Tuberculist [300] and or BioCyc [301] unless otherwise specified through its references.

3.2.3 Identification of an ESX-1 associated genes espACD binding motif

1.1) Select genes within CLR network ESX-1 related cluster:

Rv3615c; Rv3613c; Rv1639c; Rv1387; Rv3612c; Rv2406c; Rv1386; Rv1284; Rv3616c; Rv2632c; Rv2302; Rv3614c

1.2) Fusion network select genes in neighbourhood:

Rv3615c; Rv3613c; Rv1639c; Rv1387; Rv3612c; Rv2406c; Rv1386; Rv1284; Rv3616c; Rv2632c; Rv2302; Rv3614c; Rv0143c; Rv3709c; Rv1738; Rv1293; Rv1294; Rv0080; Rv0569; Rv3341; Rv2623; Rv2837c; Rv2626c; Rv0056; Rv3907c
* See homology pairs in bbh network in Table 1

1.3) New selection, select all homologs of TB31.7 in bbh network:

Rv3134c; Rv2624c; Rv2623; Rv1996; Rv2005c; Rv2026c; Rv2028c

1.4) Naïve operon extension:

Rv3134c; Rv2624c; Rv2623; Rv1996; Rv2005c; Rv2026c; Rv2028c; Rv3133c; Rv3132c; Rv3135; Rv3136; Rv3137; Rv3138; Rv3139; Rv3140; Rv3141; Rv1997; Rv2004c; Rv2003c; Rv2006; Rv2025c; Rv2024c; Rv2023c; Rv2022c; Rv2021c; Rv2020c; Rv2027c

1.5) Apply Meme2Fimo

1.6) Take top hits from Meme2Fimo result:

Rv1731; Rv3134c; Rv1997; Rv1996; Rv3135; Rv2006; Rv2623; Rv1730c; Rv2005c; Rv2023c; Rv2024c

1.7) Apply Meme2Fimo: motif D

1.8) Select genes whose upstream regions containing the motif from Meme2Fimo final selection:

gabD2; Rv1730c; PPE50; Rv3134c; otsB1; Rv2005c; Rv1996; TB31.7; ctpF; Rv2024c; Rv2023c

3.2.4 Identification of universal stress protein motif associated to ESX-1

2a.1) New selection, reselect all homologs in bbh network of TB31.7: *Rv3134c; Rv2624c; Rv2623; Rv1996; Rv2005c; Rv2026c; Rv2028c*

2a.2) Select related genes in STRING neighbourhood network:

Rv0844c; Rv3134c; Rv3133c; Rv2624c; Rv2625c; Rv2626c; Rv3132c; Rv2620c; Rv2621c; Rv2627c; Rv0845; Rv1997; Rv1996; Rv1995; Rv2619c; Rv2032; Rv2006; Rv2004c; Rv2003c; Rv1992c; Rv1993c; Rv1994c; Rv2031c; Rv2030c; Rv2622; Rv2623; Rv2029c; Rv2005c; Rv2026c; Rv2025c; Rv2028c; Rv2027c

2a.3) Select extra related genes seen by subsequent numbering (possible operon) and also co-expression in CLR network: *Rv0844c; Rv3134c; Rv3133c; Rv2624c; Rv2625c; Rv2626c; Rv3132c; Rv2620c; Rv2621c; Rv2627c; Rv0845; Rv1997; Rv1996; Rv1995; Rv2619c; Rv2032; Rv2006; Rv2004c; Rv2003c; Rv1992c; Rv1993c; Rv1994c; Rv2031c; Rv2030c; Rv2622; Rv2623; Rv2029c; Rv2005c; Rv2026c; Rv2025c; Rv2028c; Rv2027c; Rv2617c*

2a.4) Select extra related genes in STRING neighbourhood network: *Rv0844c; Rv3134c; Rv3133c; Rv2624c; Rv2625c; Rv2626c; Rv3132c; Rv2620c; Rv2621c; Rv2627c; Rv0845; Rv1997; Rv1996; Rv1995; Rv2619c; Rv2032; Rv2006; Rv2004c; Rv2003c; Rv1992c; Rv1993c; Rv1994c; Rv2031c; Rv2030c; Rv2622; Rv2623; Rv2029c; Rv2005c; Rv2026c; Rv2025c; Rv2028c; Rv2027c; Rv2617c; Rv2618c*

2a.5) Apply Meme2Fimo: motif E

2a.6) Select genes whose upstream regions containing the motif from Meme2Fimo result: *Rv2618; Rv2617c; Rv1995; Rv1994c; otsB1; Rv2005c; acg; hspX; Rv0845; narL; Rv2622; Rv2621c; ctpF*

3.2.5 Identification of universal stress protein motif associated to ESX-1 within DevR

2b.1) Select genes within CLR network ESX-1 related cluster:

Rv3615c; Rv3613c; Rv1639c; Rv1387; Rv3612c; Rv2406c; Rv1386; Rv1284; Rv3616c; Rv2632c; Rv2302; Rv3614c

2b.2) Deselect all known ESX-1 associated genes:

Rv2406c; Rv2302; Rv2632c

2b.3) Select related genes fusion network, all 3 are in the same blob:

Rv2632c; Rv2302; Rv2406c; Rv0143c; Rv2623; Rv1738; Rv2837c; Rv2626c; Rv0056; Rv0080; Rv0569; Rv3341; Rv3709c; Rv1293; Rv1294; Rv3907c

2b.4) Select related genes in STRING-neighborhood network, 3 genes are in one blob (TB31.7, RV1738, Rv0080):

Rv2632c; Rv2302; Rv2406c; Rv0143c; Rv2623; Rv1738; Rv2837c; Rv2626c; Rv0056; Rv0080; Rv0569; Rv3341; Rv3709c; Rv1293; Rv1294; Rv3907c; Rv2620c; Rv2621c; Rv0078A; Rv2619c; Rv0079; Rv2748c; Rv2751; Rv2750; Rv2622; Rv1736c; Rv1737c; Rv1734c; Rv1735c; Rv1732c; Rv1733c; Rv2749

2b.5) Select all homologous in bbh network of TB31.7:

Rv1738; Rv2632c; Rv2626c; Rv2620c; Rv2621c; Rv2302; Rv3341; Rv3709c; Rv2837c; Rv0078A; Rv2619c; Rv2406c; Rv0056; Rv0079; Rv0080; Rv3907c; Rv2748c; Rv2751; Rv2750; Rv1293; Rv1294; Rv0143c; Rv2622; Rv2623; Rv1736c; Rv1737c; Rv1734c; Rv1735c; Rv1732c; Rv1733c; Rv2749; Rv0569; Rv2026c; Rv3134c; Rv2624c; Rv1996; Rv2028c; Rv2005c

2b.6) Apply Meme2Fimo

2b.7) Take list of top hits from Meme2Fimo result:

Rv1733C; Rv0079; Rv1737C; Rv1738; Rv1996; Rv2623; Rv2005C; Rv3134C; Rv1735C; Rv1734C; Rv0569; Rv2626C; Rv1997; Rv2825C; Rv2031C; Rv2032; R v3033; Rv2338C; Rv2339; Rv0848; RV0961; Rv0574C; Rv1643; Rv0667; Rv1015C; Rv2795C; Rv1574; Rv0522; Rv1813C

2b.8) Q select (hold q key down while doing the selection) to make a sub-selection in the current selection of genes, in this case a sub-selection within the selected genes in the DevR regulon related cluster in the CLR network:

Rv1738; Rv3134c; Rv2626c; Rv0574c; Rv1997; Rv1996; Rv0079; Rv1813c; Rv2032; Rv2031c; Rv2623; Rv2005c; Rv1737c; Rv1733c; Rv0569

2b.9) Apply Meme2Fimo

2b.10) Take top hits from Meme2Fimo result: *Rv1737C; Rv1738; Rv1813C; Rv2031C; Rv2032; Rv0079; Rv1996; Rv0574C; Rv2623; Rv2005C; Rv1733C; Rv1997; Rv3134C; Rv3130C; Rv3131; Rv0848; Rv0569; Rv3409C; Rv1628C; Rv1629; Rv2626C*

2b.11) Apply Meme2Fimo: motif C

2b.12) Select genes whose upstream regions containing the motif from Meme2Fimo final selection:

Rv1738; narK2; acg; hspX; Rv3131; tgs1; Rv0079; TB31.7; Rv1813c; Rv2005c; Rv0574c; ctpF; Rv1996; cysK2; Rv3134c; Rv1733c; polA; Rv1628c

3.2.6 Identification of a likely SigE binding motif within the DevR regulon

3.1) Continue from step 3.4:

Rv2632c; Rv2302; Rv2406c; Rv0143c; Rv2623; Rv1738; Rv2837c; Rv2626c; Rv0056; Rv0080; Rv0569; Rv3341; Rv3709c; Rv1293; Rv1294; Rv3907c; Rv2620c; Rv2621c; Rv0078A; Rv2619c; Rv0079; Rv2748c; Rv2751; Rv2750; Rv2622; Rv1736c; Rv1737c; Rv1734c; Rv1735c; Rv1732c; Rv1733c; Rv2749

3.2) Q select only genes in the DevR regulon related cluster in CLR network:
Rv1738; Rv2626c; Rv0079; Rv0080; Rv2623; Rv1737c; Rv1733c; Rv0569; Rv2625c; Rv2624c; Rv0081; Rv0570

3.3) Naïve operon extend:

Rv1738; Rv2626c; Rv0079; Rv0080; Rv2623; Rv1737c; Rv1733c; Rv0569; Rv2625c; Rv2624c; Rv0081; Rv0570; Rv1736c; Rv1735c; Rv1734c; Rv1732c; Rv0082; Rv0083; Rv0084; Rv0085; Rv0086; Rv0087; Rv0088; Rv0089; Rv0090; Rv0091; Rv0092

3.4) add select in neighbourhood connect 2 large groups & remove TB31.7 as neighbours are not in DevR regulon:

Rv1738; Rv2624c; Rv2625c; Rv2626c; Rv2627c; Rv0079; Rv0081; Rv0080; Rv0083; Rv0082; Rv0086; Rv0087; Rv0084; Rv0085; Rv0088; Rv0089; Rv0090; Rv0092; Rv0091; Rv0570; Rv2628; Rv2629; Rv1736c; Rv1737c; Rv1734c; Rv1735c; Rv1732c; Rv1733c; Rv0569; Rv0567; Rv0568; Rv2631; Rv2630

3.5) Apply Meme2Fimo

3.6) Take top hits from Meme2Fimo result:

Rv0079; Rv1737C; Rv1738; Rv2031C; Rv2032; Rv1733C; Rv2627C; Rv2628; Rv1735C; Rv2629; Rv1997; Rv0089; Rv0569; Rv1734C

3.7 Apply Meme2Fimo: motif B

3.8) Select genes whose upstream regions containing the motif from Meme2Fimo final selection:

Rv0079; Rv1738; narK2; acg; hspX; Rv2628; Rv2627c; ctpF; Rv1733c; Rv2629; cysK2; Rv1735c; Rv1734c; Rv0569; Rv0080

3.3 Results

3.3.1 Identification of an ESX-1 associated genes espACD binding motif

ESX-1 related genes, *espACD*, and other closely positioned genes in the CLR network were selected. The CLR network is build based on gene co-expression data. The gene selection was transferred to the *fusion network*. The *fusion network* from STRING [302] network is based on fusion between proteins in other bacteria and co-occurrences in the same operon. Three additional genes were identified in their neighbourhood. This selection was further enlarged with genes in their neighbourhood previously reported in the DevR regulon [52]. Transferring the selection to the bbh network, that contain operon and BLAST based homologs, led to the identification of three pairs of homologous genes. In each pair one gene belongs to the ESX-1 related gene set whereas the other one is in the DevR regulon and are shown in **Table 2**.

Table 2. Hypothetical homologous complex.

ESX-1 CLUSTER RELATED	DEVR CLUSTER RELATED
RV0569	*Rv2302*
RV2632C	*Rv1738*
RV2406C*	*Rv2626c**
	Rv0080
	TB31.7

*Pairs of homolog genes in the ESX-1 and DevR related clusters of two hypothetical homologous complexes. * low similarity (E-value 3e-09 < network visualization threshold)*

In the fusion network, genes in these homology pairs within the DevR regulon appear as a densely connected cluster, together with Rv0080 and TB31.7. TB31.7 is a universal stress protein family protein responding to stress signals, interacts with cAMP [303] and has been shown to be involved in growth arrest during latent infection [304], [305]. To further investigate the role of TB31.7 a new selection was made in the bbh network by adding six TB31.7 homologs, five of which are in the DevR regulon. Meme2Fimo was iteratively used to explore upstream sequences of these genes. Finally, a conserved motif similar to the one reported for DevR was identified and is shown in **Figure 7**.

However, some distinct features appear showing that regulation of these ESX-1 related genes is complex, integrating signals from hypoxia via DevR as well as signals from cell stress signals via TB31.7 homologs.

Similarities and differences between the various motifs will be discussed in more detail in the section “Motif comparison”. The final list of genes used to build this motif can be found in **Table 3**.

Table 3. Final selection of genes used to build motif D of ESX-1, espACD associated stress response.

Locus tag or name	Functional information
gabD2	Putative succinate-semialdehyde dehydrogenase
Rv1730c	Possible penicillin-binding protein
PPE50	Suspected to be involved in epitope variation [306]
Rv3134c	Universal stress protein family protein, part of DevR-DevS operon
otsB1	trehalose-phosphatase
Rv2005c	Universal stress protein family protein
Rv1996	Universal stress protein family protein, predicted possible vaccine candidate (See Zvi et al., 2008).
TB31.7	a universal stress protein family protein involved in growth arrest
ctpF	probable metal cation transporter P-type ATPase A (CtpF)
Rv2024c	Conserved hypothetical protein
Rv2023c	NA

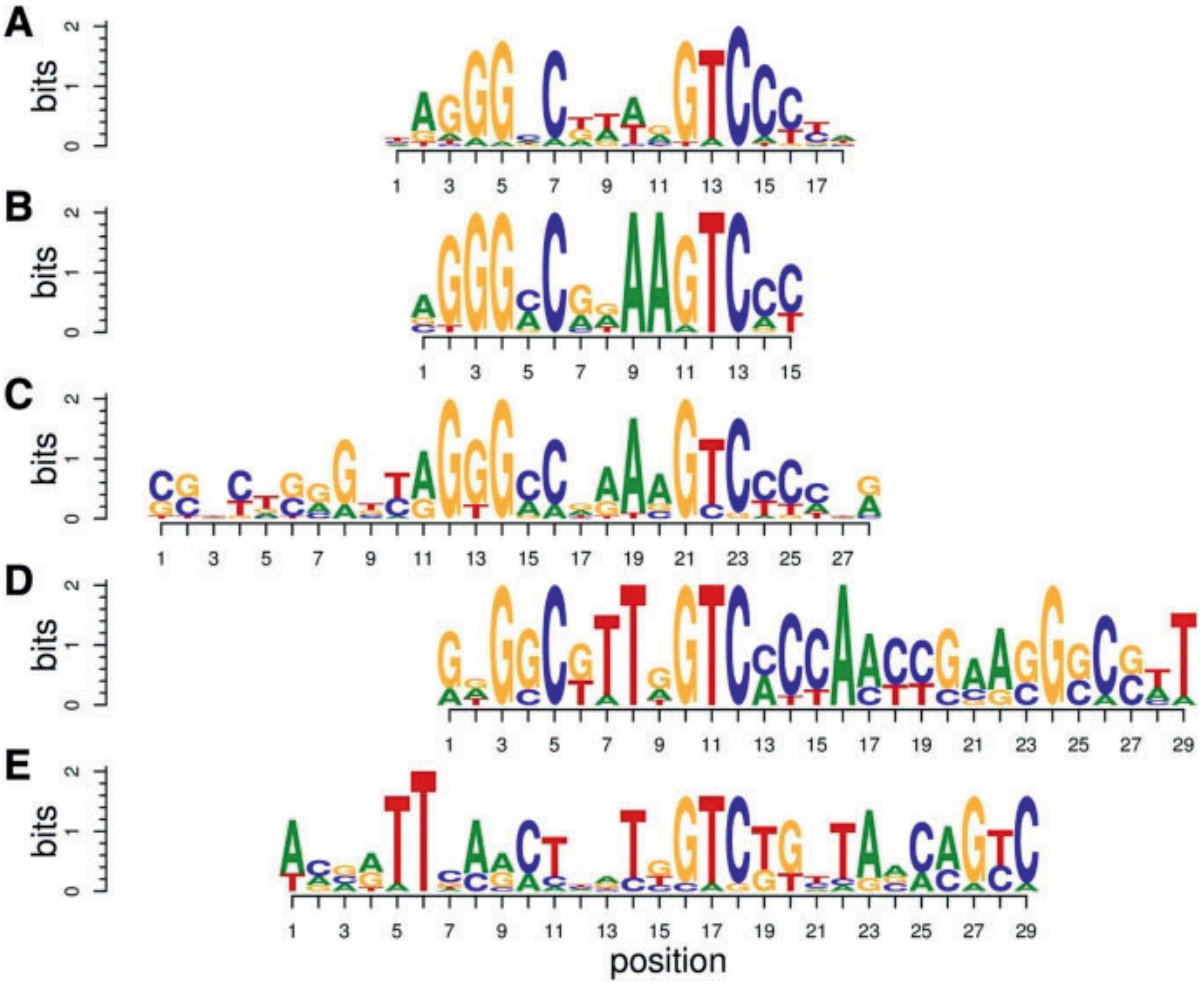

Figure 7. Comparison of DevR and ESX-1 related motifs. ***a*** *DevR motif as reported in* [242] (***b***) *Exploration path 3 motif.* ***c*** *Exploration path 2 motif 2.* ***d*** *Exploration path 1 motif.* ***e*** *Exploration path 2 motif 1. See Fig. 7 for the legend (E. Lindfors et al.*[51]*).*

Most of the genes in the final cluster are associated to response to against stress in the phagosome. For example, *Rv1730c* is possibly penicillin binding, *Rv3134c* is part of the *DevR-Devs* operon and is involved in phosphor relay of DevR, *otsB1* is involved in Mycolic acid synthesis, *Rv2005c* Is linked to phenotypic fluoroquinolone resistance [307], *Rv1996*is linked to zinc efflux to protect from zinc poisoning, *TB31.7* codes for a universal stress protein, *Rv2024c* is mamB DNA methylation [308]. *EspACD* is a highly regulated virulence operon, see **Figure 8.**

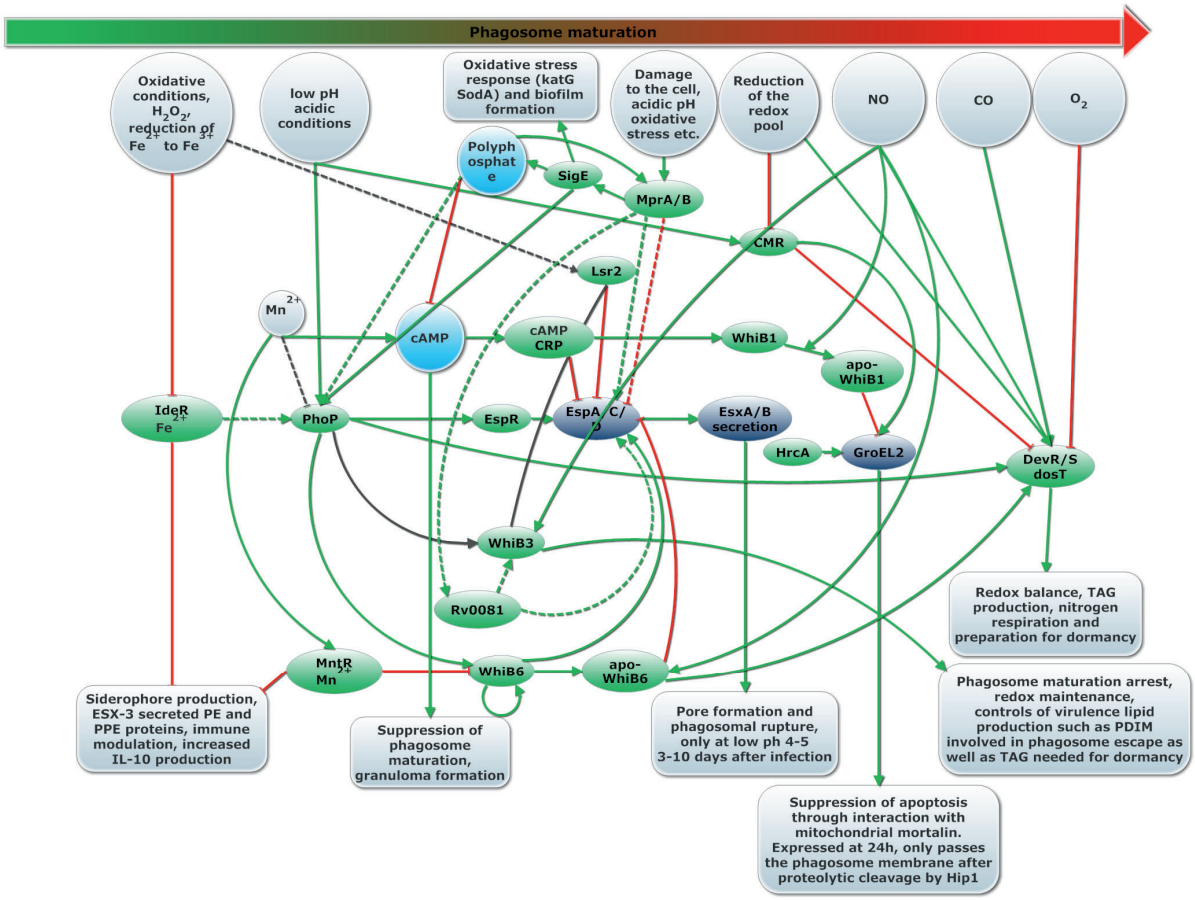

Figure 8. Overview of the regulatory cascade that integrates environmental cues to active the immune modulation, dormancy, and phagosome escape virulence strategies. Arrows indicated regulation (green for induction, red for inhibition of transcription) with dashed lines. This figure was earlier presented as Figure 4 in chapter 1.

We see that the homologs and similarly regulated gene we identified are all involved in adaptation to intraphagosomally stress conditions such as hypoxia, antibiotics of which some are known to be regulated by PhoP and link to preparation for dormancy. These findings are in line with what know about the regulation of *EspACD*, which we identify as a key hub in regulation of virulence and phagosomal pore formation and subsequent escape as well as preparation for dormancy. The unknown regulator of these ESX-1 associated genes would be in interesting drug targets, since it makes *Mtb* unable to escape and more vulnerable to drugs and the immune system.

3.3.2 Identification of universal stress protein motif associated to ESX-1

To further investigate the *TB31.7* universal stress protein gene and its homologs, we selected them and neighbouring genes within the neighbourhood network. Upstream regulatory regions analysis led to the description of another motif **Figure 7**, Motif E. A subset of genes (*Rv2621c*, *Rv2622*), coding for a possible transcriptional regulator and methyltransferase, with this motif in their upstream regions appear in the CLR network with a cluster of genes related to mycolic acid synthesis (**Table 4**). The ratio of free and bound mycolic acids is known to change under hypoxia and cell wall stress

[45]. We further investigated *TB31.7* homologs that are part of the DevR regulon and identified motif (**Table 5**).

Table 4. Final selection of genes used to build motif D, TB31.7 associated stress proteins.

Locus tag or name	Functional information
Rv2618	Conserved hypothetical protein
Rv2617c	Probable transmembrane protein
Rv1995	Unknown protein
Rv1994c	Metal sensor transcriptional regulator CmtR (ArsR-SmtB family)
otsB1	TPP; trehalose-phosphatase
Rv2005c	Universal stress protein family protein. Predicted possible vaccine candidate (See Zvi et al., 2008).
acg	Putative NAD(P)H nitroreductase acg
hspX	HspX promotes the polar localization of mycobacterial protein aggregates
Rv0845	Possible two component sensor kinase
narL	NarL, nitrate/nitrite response regulator protein
Rv2622	Possible methyltransferase (methylase)
Rv2621c	Possible transcriptional regulatory protein
ctpF	probable metal cation transporter P-type ATPase A (CtpF)

Table 5. Final selection of genes used to build motif E, TB31.7 associated stress proteins within the DevR regulon, Motif E.

Locus tag or name	Functional information
Rv1738	Conserved protein, implicated in the onset of nonreplicating persistence [309]
narK2	*narK2* encodes a nitrate, H^+ symporter
acg	Putative NAD(P)H nitroreductase acg
hspX	HspX promotes the polar localization of mycobacterial protein aggregates
Rv3131	Conserved protein
tgs1	NA
Rv0079	Part of DevR regulon, appears to be involved in the regulation of translation through the interaction of its product with bacterial ribosomal subunits [309]
TB31.7	NA
Rv1813c	Conserved hypothetical protein
Rv2005c	Universal stress protein family protein
Rv0574c	Conserved hypothetical protein
ctpF	probable metal cation transporter P-type ATPase A (CtpF)
Rv1996	Universal stress protein family protein

cysK2	O-Phospho-l-Serine-Dependent S-Sulfocysteine Synthase
Rv3134c	Universal stress protein family protein
Rv1733c	Probable conserved transmembrane protein
polA	DNA polymerase I, DNA repair
Rv1628c	Conserved protein

3.3.3 Identification of a likely SigE binding motif within the DevR regulon

We explored the DevR regulon to identify elements with additional regulatory influences. USPs homologs to *TB31.7* with the DevR regulon and genes in the same operons were selected. Transferring the selection to the gene neighborhood network showed the relationship between these two related groups and suggested some genes to be further included in the selection. Yet another motif (**Figure 2**,**Table 6**) was described in the upstream regions of these genes. This motif is similar to the binding motif of the AlgU sigma factor from *P. aeruginosa* which is homologous to SigE in *M. tuberculosis* [310]. SigE and SigH together with MprAB function to detect and protect against cell stress such as misfolded proteins, heat shock, acidic pH, exposure to detergent, and oxidative stress. These conditions are associated with failed immune modulation which is related to the DevR regulated dormancy regulon [310], [311]. Moreover, *Rv0080*, which is also in the DevR regulon, has been reported as a regulatory hub of the hypoxia response regulated by MprA [45], [269]. The identified binding motif shows similarity to the motifs detected upstream of genes experimentally shown to be regulated by SigE and SigH regulated genes [51], *NarK2* is involved in nitrate expulsion, *Tgs1* is involved in triacylglycerol production, *Rv0574* is a possible polyglutamate synthase involved in encapsulation and *cysK2* is a probable cysteine synthetase [300]. Most of the genes identified in the SigE regulon are of unknown or hypothetical function. However, for many it is known they are expressed in hypoxia and dormancy conditions and some like *hspX* are known to be directly DevR regulated [300].

Table 6. Final selection of genes used to identify the likely SigE binding motif.

Locus tag or name	Functional information
Rv0079	Part of DevR regulon, appears to be involved in the regulation of translation through the interaction of its product with bacterial ribosomal subunits [309]
Rv1738	Conserved protein, implicated in the onset of nonreplicating persistence [309]
narK2	*narK2* encodes a nitrate, H^+ symporter
acg	Putative NAD(P)H nitroreductase acg
hspX	HspX promotes the polar localization of mycobacterial protein aggregates
Rv2628	Hypothetical protein, associated with latent tuberculosis infection [309]
Rv2627c	Conserved protein
ctpF	probable metal cation transporter P-type ATPase A (CtpF)

Rv1733c	Probable conserved transmembrane protein
Rv2629	Conserved protein
cysK2	NA
Rv1735c	Hypothetical membrane protein, part of DevR
Rv1734c	Conserved hypothetical protein. Similar to Acetyltransferase from Chlamydia pneumoniae
Rv0569	Conserved protein
Rv0080	Homolog of Rv2406c which contains a cAMP and CRP-binding element

3.3.4 Motif comparison

Figure 9 shows five related binding motifs. The location of these motifs is shown in **Figure 10**. The groups of genes controlled by these motifs are shared as shown in Figure 5. Inspection of the locations of the motifs shows their overlaps in the upstream regions of the various shared genes of motifs B, C and D, which indicates that the shifted motifs might still be functional. The general DevR motif GGGNCNNNNGNCCC is palindromic, whereas motif B GGGNCNNAAGTC has a unique element, which is not palindromic. Both SigE and DevR are related to the modulation of process directly related to growth within human macrophages, the similarity between this motif and the AlgU motif in *P. aeruginosa* led us to hypothesize that DevR and SigE can bind to the same regions. Furthermore, motif D GGGNCNTTNGTC also has a unique element, NAA in motif B is replaced by TTN. The palindromic motif E lacks the characteristic GGGNCNNNNGNCCC pattern describing the general DevR binding motif. Only the GTC is conserved in comparison to the other motifs. The regions it matches are close (14 and 37 nucleotides) to the regions matched by motif B. Therefore, we hypothesize that this motif might be associated to additional regulatory elements.

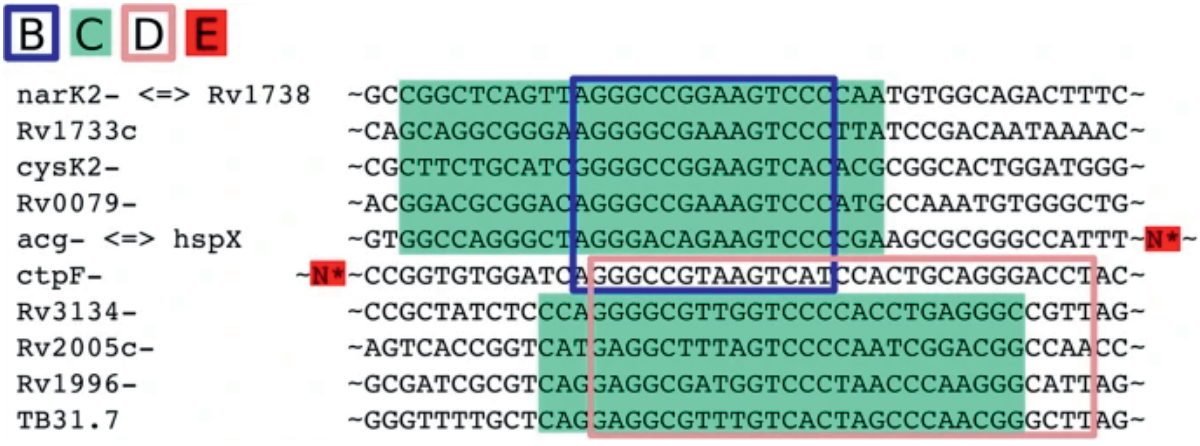

Figure 9. Shifted motif alignment. Marked region denotes the region containing the sequence to which the motif matches. The regions marked for the motif D regions are shifted. See Fig. 7 for the legend (E. Lindfors et al.[51]*)*

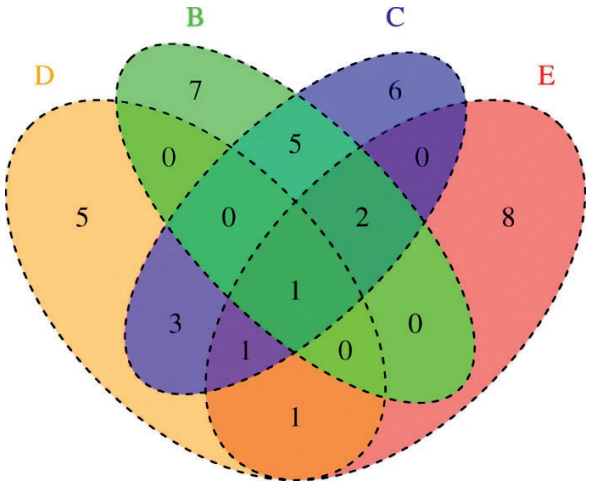

Figure 10. Shared genes. Presence of binding motifs A, B, C and D in gene upstream regions. See Fig. 7 for Legends A, B, C and D motif description See Fig. 9 for the legend (E. Lindfors et al. [51])

3.4 Discussion

We have shown how SyNDI can be used to explore and better understand complex regulated systems such as ESX-1 and associated virulence proteins in *Mtb*. In addition, we were able to detect multiple and related binding motifs within the DevR regulon which have not yet been described in the literature, including a motif that we hypothesize it is related to *Mtb* SigE. Identification of regulatory motifs is not only important to increase our understanding of regulation, but also to identify potential drug and knock out (KO) targets for vaccine development. In exploration path 1, we identified a regulatory binding site of which its unknown regulator would be an interesting drug target based on the multiple genes in this regulon being involved to adapting to hypoxic and cellular stress as well as two genes involved in antibiotic resistance and one gene known to protect against metal poisoning. Similarly, we identified a binding motif of SigE which is involved in the stress response which we show, overlaps with the DevR, dormancy regulon. SigE is needed to arrest phagosome maturation [312]. KO mutants of SigE show increased susceptibility to heat and chemical stress [290]. SigE KO mutants were shown to be effective vaccine candidates in guinea pig [279] and to be even more effective as vaccine candidate in combination with *faD26* KO in a mouse model [313]. Similarly, combined faD26 and PhoP KO mutants were shown to be an effective in vaccination, although the lack of a functioning ESX-1 system reduced the immunity against wild type *Mtb* [290]. Based on our results, it can be easily understood why SigE is a potential vaccine or drug targets.

As we can see in the alignment of the four identified regulatory motifs, they overlap. This overlap indicates regulation of these virulence associated genes is likely to be complex involving multiple regulators with overlapping binding sites, as well as complex regulation through cooperative binding of DevR to multiple upstream binding [241], [314] and cooperative binding of MprAB to multiple binding sites [315].

MprAB regulates SigB and SigE involved in the stress response [315] while PepD [311]. and RspA [316] regulate SigE which according to our findings regulates some genes present in the DevR regulon. Only by large scale mapping of all these regulators, their regulatory binding motifs, and the genes they regulate, can we hope to understand the complexity of regulation in *Mtb*.

The Meme2Fimo method which we used to unravel gene regulation in *Mtb,* illustrates a few common elements of the data-drive hypothesis generation research performed in in Systems Biology: the use of a) (multi-)omics data to identify clusters of interest, b) iterative approaches to identify and refine clusters and remove noise and c) Integration with existing knowledge from online resources and literature to contextualize the findings. Here, the exploration started with the question "*are there sub clusters of genes in the SigE and DevR regulon regulated by other regulatory proteins?*" Integrative solutions such as Diva and SyNDI are important because they allow researchers to explore and identify clusters and motifs and to answer biological questions in a reasonable amount time. The Meme2Femo tool used for segregating motifs was effective, however it still required manual input from the user to select and deselect genes for each iteration. Full automation of the exact workflow involving manual selections in multiple networks is not yet possible. Users can make intelligent decisions over multiple network which a computer algorithm cannot mimic. However, it is very well possible to perform automatic cluster detection and as such automatic identification of motifs in a selected network. The most logical choice for automated cluster and motif detection would be automatic cluster detection in networks such as the CLR network which we used extensively in all discussed explorations. Better integration of automated cluster and motif detection with manual curation could be a topic for future research.

Chapter 4
Phenotype and multi-omics comparison of *Staphylococcus* and *Streptococcus* uncovers pathogenic traits and predicts zoonotic potential

Adapted from:

Niels A. Zondervan, Vitor A. P. Martins dos Santos, Maria Suarez-Diez, Edoardo Saccenti. "Phenotype and multi-omics comparison of *Staphylococcus* and *Streptococcus* uncovers pathogenic traits and predicts zoonotic potential". In *BMC genomics* 22(1) 2021.

4.1 Abstract

Background

Staphylococcus and *Streptococcus* species can cause many different diseases, ranging from mild skin infections to life-threatening necrotizing fasciitis. Both genera consist of commensal species that colonize the skin and nose of humans and animals, and of which some can display a pathogenic phenotype.

Results

We compared 235 *Staphylococcus* and 315 *Streptococcus* genomes based on their protein domain content. We show the relationships between protein persistence and essentiality by integrating essentiality predictions from two metabolic models and essentiality measurements from six large-scale transposon mutagenesis experiments. We identified clusters of strains within species based on proteins associated to similar biological processes. We built Random Forest classifiers that predicted the zoonotic potential. Furthermore, we identified shared attributes between of *Staphylococcus aureus* and *Streptococcus pyogenes* that allow them to cause necrotizing fasciitis.

Conclusions

Differences observed in clustering of strains based on functional groups of proteins correlate with phenotypes such as host tropism, capability to infect multiple hosts and drug resistance. Our method provides a solid basis towards large-scale prediction of phenotypes based on genomic information.

4.2 Background

Species from the genera *Staphylococcus* and *Streptococcus* are mostly commensals that live as part of the microbiota of various animals and humans [317]. Some of them are opportunistic pathogens, displaying a pathogenic phenotype when the immune system of the host is compromised or the epithelial barrier is damaged [318]–[321].

Few comparative genomic studies have been performed to analyse the evolution and the pathogenesis of *Staphylococcus* and *Streptococcus* species: the comparisons of the genomes of 11 *Staphylococcus* species determined that horizontal gene transfer of virulence factors is an important factor in adaptation of *S. aureus* to humans [322]; another study showed that protein domain based metabolic diversity among *Streptococcus* species could be used to identify differences in the metabolism of the highly pathogenic serotype 2 *S. suis* compared to other *Streptococci* [323]. Another study confirmed these results and showed that metabolic capability predicted using genome scale models (GEMs) could be used to identify *Streptococcus* strain specific biomarkers and metabolic determinants of virulence [324].

Protein domains and protein-domain architectures have been shown to be a fast and efficient method to define groups of functionally equivalent proteins that were used for comparative genomic studies [325], [326], including *Staphylococcus* and *Streptococcus* [327]–[329]. However, at the best of our knowledge, no work exists focusing on similarities and differences within and between *Staphylococcus* and *Streptococcus* genomes.

In this study we performed a comparative analysis of 235 and 315 fully sequenced *Staphylococci* and *Streptococci* genomes by annotating their proteins based on their domain content. We integrated this protein annotation with genome-scale metabolic-modelling predictions, transcriptomic and transposon-mutagenesis data sets to study gene essentiality and persistence. All annotation used in this paper as well as GO information is based on genomics annotation from databases based mainly on bacterial genomics studies. In this paper we compare within and between *Staphylococcus* and *Streptococcus* species with the objective to identify both difference and similarities in genomic properties as well as in specific combinations of genes that give rise to pathogenic phenotypes. We compared the clustering of *Staphylococcus* and *Streptococcus* genomes based on proteins selected using on Gene Ontology (GO) terms associated with clinical phenotypes such drug resistance, pathogenesis, and tissue and host tropism. Furthermore, we used the functional grouping of proteins to predict zoonotic potential of *S. suis* and *S. agalactiae*, that is their ability to infect multiple hosts including humans. Finally, we compared *S. aureus* and *S. pyogenes* to identify the genomic basis for their shared ability to cause severe bacterial infections like necrotizing fasciitis. Our results are compared throughout the paper with findings from literature.

4.3 Results

4.3.1 Pan- and core genome analysis

The size of the pan- and core genomes of *Staphylococcus* and *Streptococcus* was determined based on protein domain content (**Figure 11**). The pangenome contains all proteins present in the analysed genomes. The core genome contains only proteins that are present in all genomes and represents their genomic essence [330]. The ratio of the sizes of the core- and pan genome are 0.22 (557/2563) for *Staphylococcus* and 0.17 (458/2725) for *Streptococcus*. A Heaps' regression model was used to estimate the closedness of the pangenome [331]. The closedness of the pangenome represents how much the addition of more genome sequences is expected to increase the number of proteins in the pangenome. For both *Staphylococcus* (α = 1.10± 0.02) and *Streptococcus* (α = 1.12± 0.01) the pangenome was found to be closed (*i.e* few new genes are added as news strains are discovered/sequenced). Additional plots of the estimated pan- and core genomes size and the Heaps' regression model can be found in supplementary material [see Additional file 5].

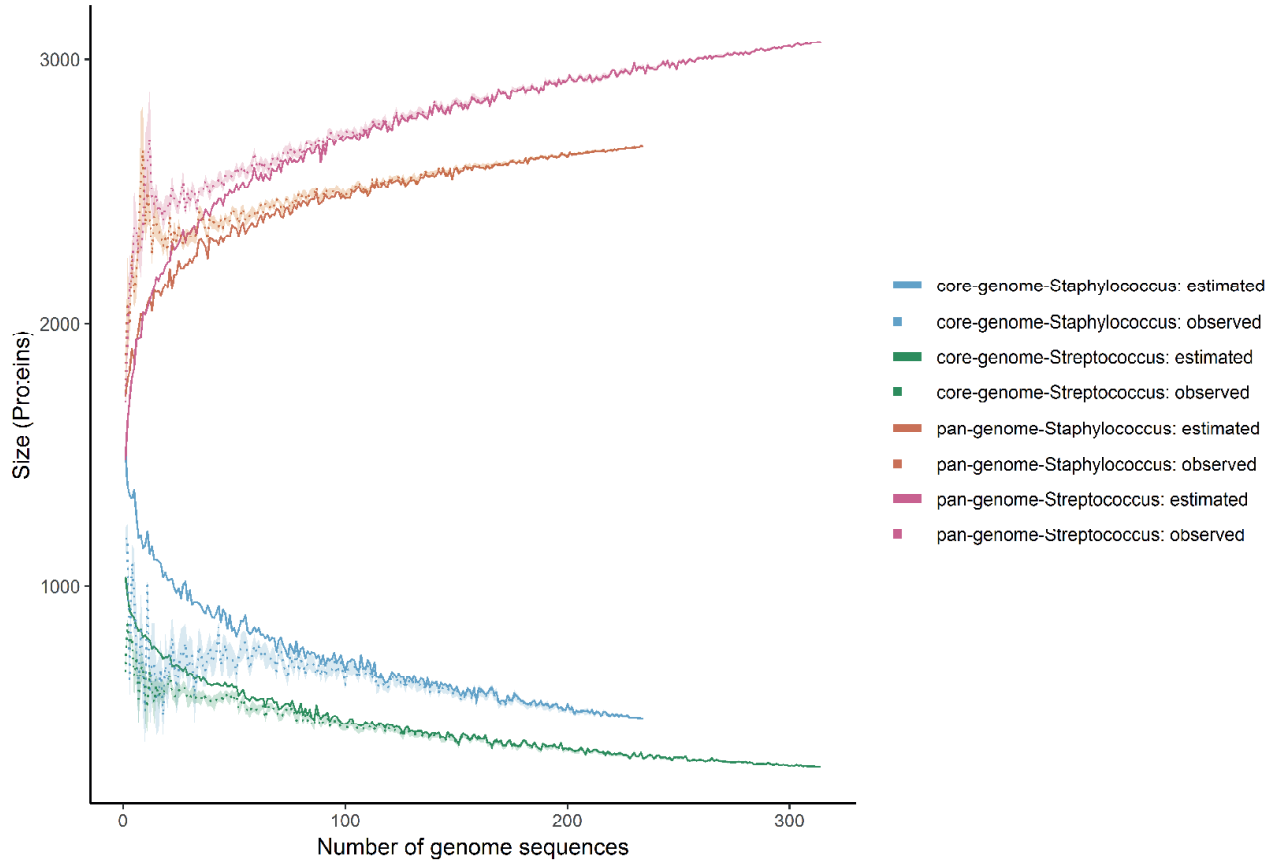

Figure 11. Mean observed and estimated size of the pan- and core genome. The shadowed area shows variation over 10 times sampling.

4.3.2 Protein persistence and essentiality

Persistence of proteins over all *Staphylococcus* and all *Streptococcus* species was calculated. Protein persistence data was combined with model predictions of essentiality and experimentally determined essentiality data. Experimentally determined essentiality (labelled as EXP) is available for growth on rich media resembling in vivo conditions. GEMs predictions were made using minimal media conditions for all combinations of carbon, nitrogen, sulphur, and phosphorus sources. Simulations on rich media conditions were therefore indirectly performed since all rich media compounds are present in the models as exchange reactions and all combinations of these exchange reactions functioning as carbon, nitrogen, sulphur, and phosphor sources were tested for essentiality. We used GEM to predict gene essentiality for *Staphylococcus aureus* NTCTC 8325 and *Streptococcus pyogenes* M49. The total number of medium combinations based on C, N, S, P sources was 12432 for *Staphylococcus* and 714 for *Streptococcus*. The number of tested conditions for Staphylococcus is much larger than the Streptococcus model since the Staphylococcus model can use all amino acids as alternative nitrogen source through deamination, greatly increasing the number of minimal media combinations. The Staphylococcus model can use all amino acids as alternative nitrogen source through deamination, greatly increasing the number of minimal media combinations" Protein persistence, *in silico* predictions of essential and *in vitro* essentiality data for *Staphylococcus* and *Streptococcus* were integrated based on their associated locus tags. Both GEM based and experimentally determined essentiality correlated with a high persistence, while essentiality by both criteria is associated with an even higher persistence (see **Table 7** and **Figure 12**). Proteins experimentally determined or GEM predicted to be essentiality are significantly different from the average protein persistence (Student's *t*-test, *p*-value = 5×10^{-14}) for both *Staphylococcus* and *Streptococcus*.

Table 7. Persistence of Staphylococcus (Staph.) and Streptococcus (Strep.) for all proteins, proteins associated to Genome Metabolic model (GEM) essential genes and experimentally (EXP) determined essential genes.

Group	Avg persistence Staph.	Avg persistence Strep.
All	0.60 ± 0.44 (*N*=2655)	0.42 ± 0.42 (*N*=3047)
GEM-essential	0.94 ± 0.14 (*N*=153)	0.98 ± 0.09 (*N*=225)
Exp-Essential	0.97 ± 0.03 (*N*=411)	0.97 ± 0.12 (*N*=254)
EXP&GEM-Essential	0.94 ± 0.01 (*N*=46)	0.98 ± 0.00 (*N*=113)

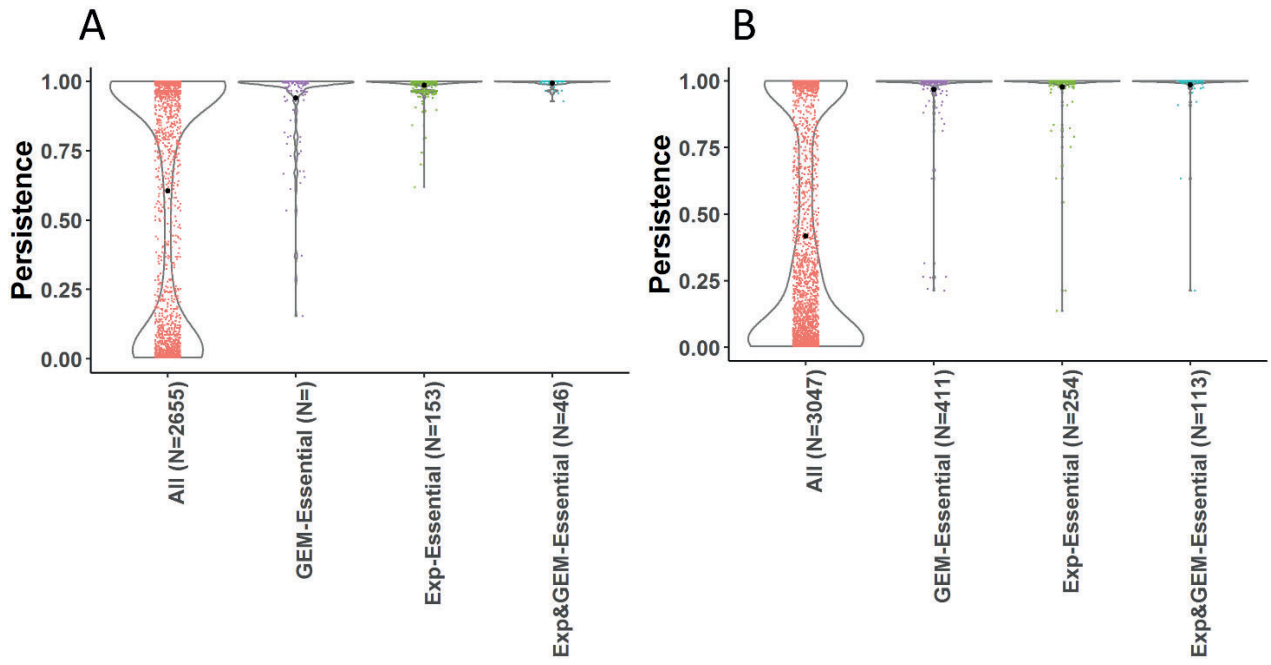

Figure 12. Protein persistence. A) Staphylococcus, B) Streptococcus. Group labels: All= all proteins, GEM = in silico predicted to be essential using a Genome Scale Metabolic model, Exp-Essential = experimentally determined to be essential. Combined group strains

4

4.3.3 Variability of gene expression and gene essentiality

Essentiality and domain persistence information for *Staphylococcus* was combined with the variability of transcription (measured by $\log_2$ fold changes). The variability in expression for experimentally determined essential and non-essential genes as well as for persistent and non-persistent genes were compared (**Figure 13**). The fold change transcription levels of experimentally determined essential genes are significantly less variable than the transcription levels of non-essential genes (Student's *t*-test, *p*-value = 5×10^{-14}) as well as for persistent genes (Student's *t*-test, *p*-value = 0.000124)).

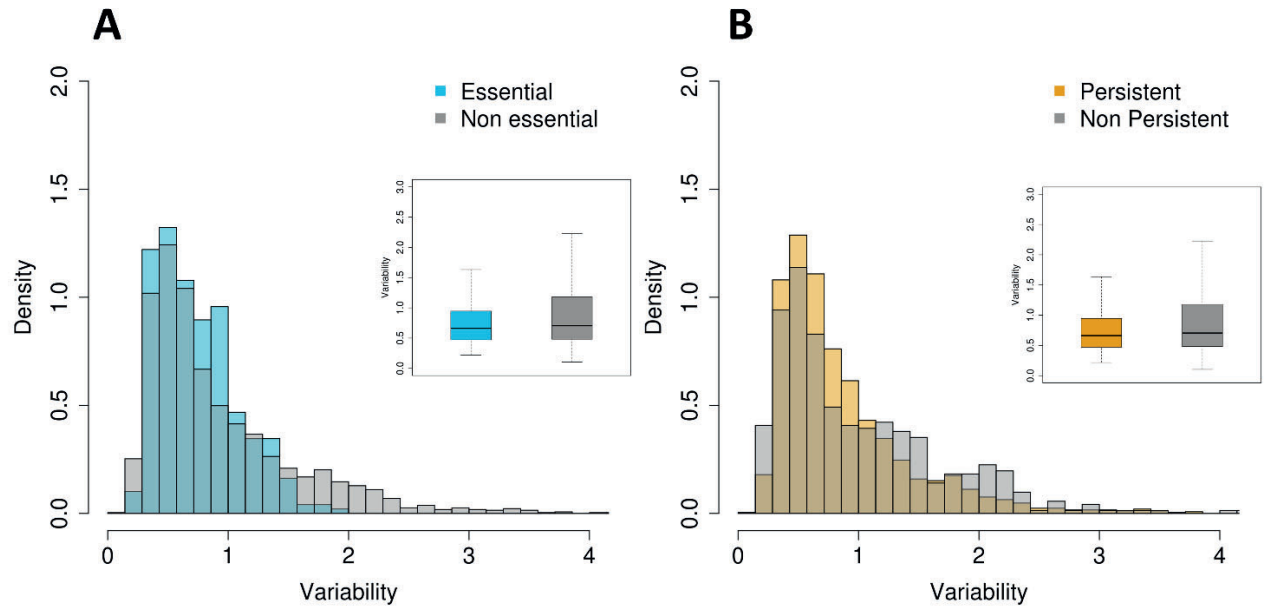

Figure 13. Transcriptional variability of essential and non-essential genes in Staphylococcus. *Box plots show Variability values for both groups. Difference between mean values is significant (p-val < 0.01). (B) Transcriptional variability of persistent and non-persistent genes (genes with persistence lower or higher than 0.95, respectively). Box plots show Variability values for both groups. Difference between mean values is significant (p-val < 0.01)*

4.3.4 Functional analysis of pathogenesis and pathogenicity

For this analysis, we filtered proteins from *Staphylococcus* and *Streptococcus* on their association to 17 genome ontology (GO) biological process terms associated to pathogenesis and pathogenicity identified from literature (**Table 8**). Filtering included all proteins associated to either the 17 main GO terms or any of their descendent terms. For all GO terms, proteins were found in both *Staphylococcus* and *Streptococcus* (**Table 9**). The ratio of proteins per GO function to the total number of proteins is similar for *Staphylococcus* and *Streptococcus* except for the group 'Biological adhesions' which has a larger fraction of proteins associated in *Streptococcus* than in *Staphylococcus*.

Table 8. Gene Ontology (GO) terms used to select proteins based on their domain content for functional trees, PCA and t-SNA analysis. GO terms that are direct children of the 'Biological process' GO term are marked with an asterisk (*).

GO ID	DESCRIPTION
GO:0008150	Biological process
GO:0008152	*Metabolic process
GO:0017144	Drug metabolic process
GO:0042493	Response to drug
GO:0023052	*Signalling
GO:0065007	*Biological regulation
GO:0022610	*Biological adhesion

GO:0044406	Adhesion of symbiont to host
GO:0051704	Multi-organism process
GO:0044419	Inter species interaction between organisms
GO:0042710	Biofilm formation
GO:0098743	Cell aggregation
GO:0044403	Symbiont process
GO:0009372	Quorum sensing
GO:0035821	Modification of morphology or physiology of other organism
GO:0009405	Pathogenesis

Table 9. Number of proteins in the pangenome of Staphylococcus and the pangenome of Streptococcus per GO term. Root ontology terms, terms without a parent, are marked in their description with an asterixis (*). GO terms are order as such that descendent GO term.

FILTER	DESCRIPTION	STAPH	STREP
	All proteins	2655	3047
GO:0008150	Biological process	1974	2222
GO:0008152	*Metabolic process	1661	1871
GO:0017144	Drug metabolic process	59	77
GO:0042493	Response to drug	56	60
GO:0023052	*Signalling	217	280
GO:0065007	*Biological regulation	823	929
GO:0022610	*Biological adhesion	70	147
GO:0044406	Adhesion of symbiont to host	2	3
GO:0051704	Multi-organism process	348	456
GO:0044419	Inter species interaction between organisms	210	309
GO:0042710	Biofilm formation	9	10
GO:0098743	Cell aggregation	7	10
GO:0044403	Symbiont process	150	202
GO:0009372	Quorum sensing	7	7
GO:0035821	Modification of morphology or physiology of other organism	45	74
GO:0009405	Pathogenesis	65	115

Functional trees, PCA and t-SNE plots were used to compare the (dis-)similarity in clustering of the genomes based on functional groups of proteins compared to clustering based on all proteins. Dissimilarity was calculated using the Euclidean distances of genomes in the functional trees and by scaling these distances to values between 0 and 1 to make them comparable. Functional trees, PCA plots and t-SNE plots for *Staphylococcus* can be found in supplementary material [Additional file 6,7 and 8]. Functional trees, PCA plots and t-SNE pots for *Streptococcus* can be found in supplementary material [Additional file 9, 10 and 11]. PCA plots and t-SNE plots for *Staphylococcus* and *Streptococcus* species combined can be found in supplementary material [Additional file 12 and 13].

Heatmaps were used to investigate which proteins are absent for each species. Each of these analyses and visualization methods has their own strength and weaknesses in showing the differences in clustering. In the following we highlight some of the differences in clustering of *Staphylococcus* and *Streptococcus* genomes based on proteins annotated per GO term as compared to clustering based on all proteins.

4.3.5 Correlation between GO functional groups of proteins

We calculated the correlation between functional trees to compare the similarity in clustering per GO functional group of proteins **(Figure 14 A-B**). The correlation between functional trees is higher for children and parent GO terms as well as for GO terms with similar functions such as 'drug metabolic process 'and 'response to drug'. In general, we see that functional trees based on fewer proteins have a lower correlation than functional trees based on many proteins. These results were expected since fewer proteins means less information to separate strains resulting in merging of branches in the tree. An interesting exception to this rule is the 'symbiont process' functional tree which has the lowest correlation with other functional trees for *Staphylococcus* even though there is a high number of proteins associated to this GO term.

There are some notable differences when comparing the correlation between functional trees for *Staphylococcus* and *Streptococcus*. For *Staphylococcus*, the 'pathogenesis' clusters together with the functional tree 'modification of morphology or physiology of other organisms'. For *Streptococcus,* the functional tree of 'pathogenesis' clusters together with the functional tree of 'biological adhesion'. Many 'modification of host morphology' proteins in *Staphylococcus* are also associated to the GO term 'pathogenesis' while many 'biological adhesion' proteins in *Streptococcus* are associated to the GO term 'pathogenesis'. These results could indicate that modification of host morphology is important for the pathology of *Staphylococcus* strains while biological adhesion is more important for the pathology of *Streptococcus*.

4.3.6 Horizontal gene transfer of proteins related to pathogenesis

The PCA plot based on all proteins combining *Staphylococcus* and *Streptococcus* genomes supplementary material [Additional file 12 and 13], shows genomes of the same species to cluster together as we would expect (**Figure 15-A**). The PCA plot based on presence/absence of proteins involved in Response to drug (GO:0042493) shows genomes are not always separated on the species level, however, there is a clear separation between *Staphylococcus* and *Streptococcus* genomes (**Figure 15-B**). However, both in the PCA (**Figure 15-C**) and in t-SNE plots based on proteins associated to the GO term 'Pathogenesis' proteins, *Staphylococcus* and *Streptococcus* species cluster together.

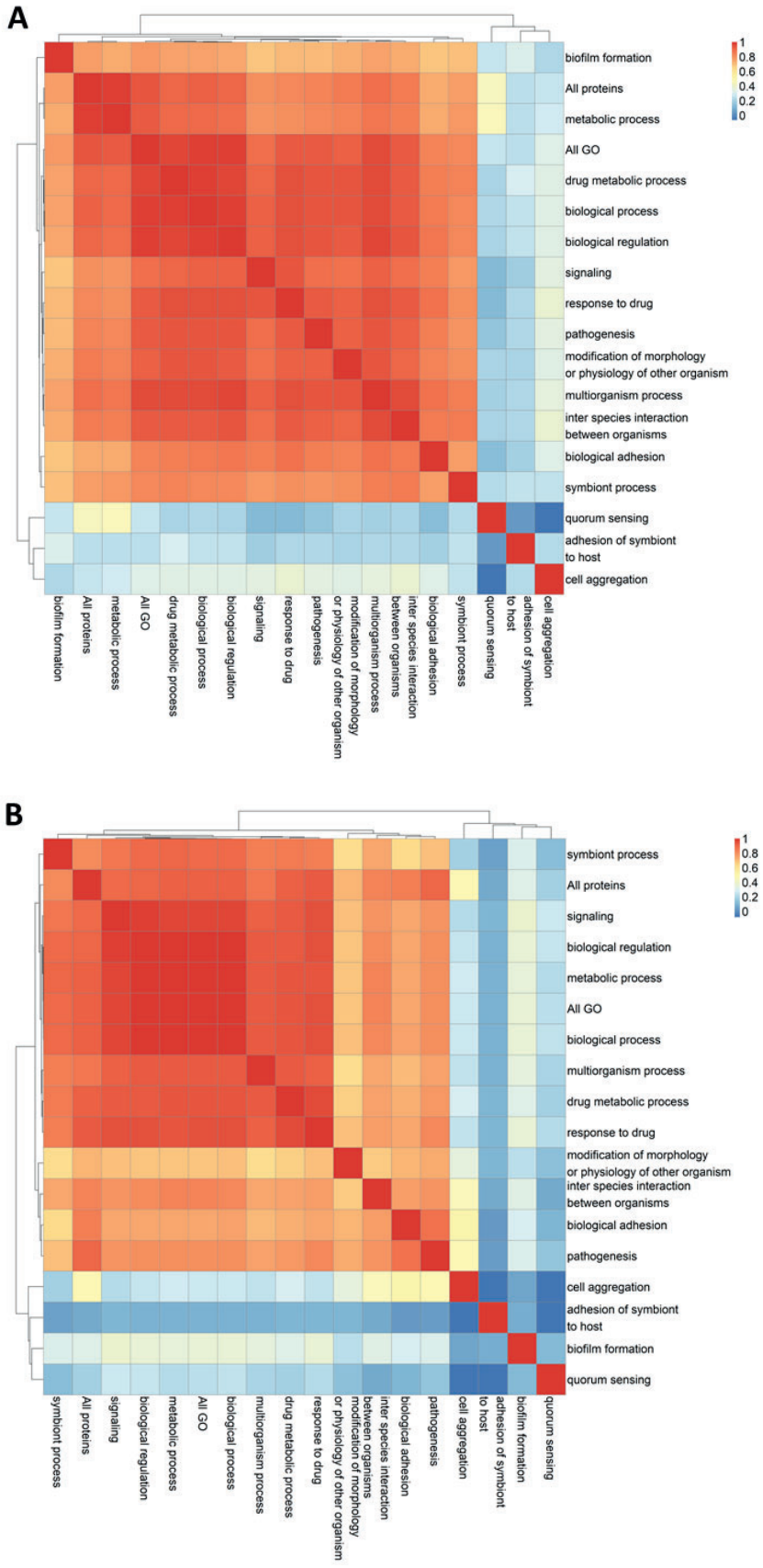

Figure 14. A) Heatmaps of the correlation between Staphylococcus *functional trees, B) Heatmaps of the correlation between* Streptococcus *functional trees.*

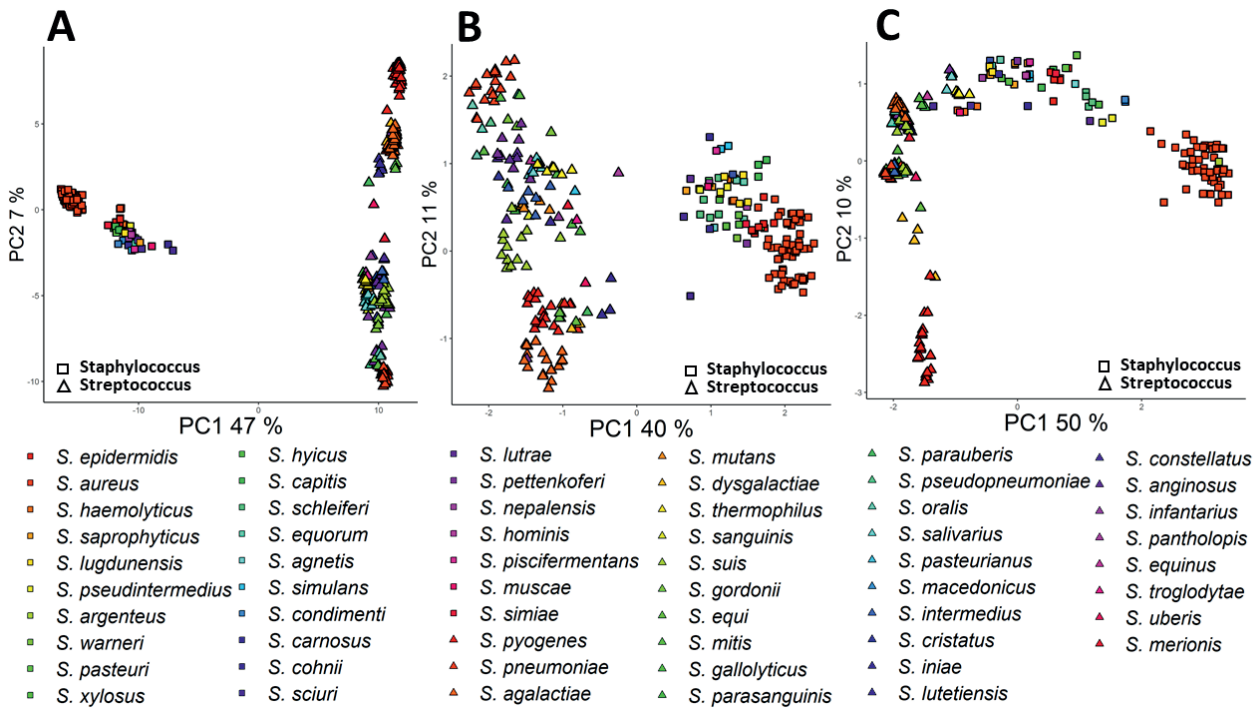

Figure 15. (A) PCA based on all proteins in Staphylococcus *and* Streptococcus. *(B) PCA based on proteins association to 'response to drug' (GO:0042493). (C) PCA based on proteins associated to 'pathogenesis' (GO:009405). Fraction of variance explained by each PC is indicated in the axis*

Analysis of the presence and absence of proteins associated to 'pathogenesis' reveals that *Staphylococcus sciuri* GCA:002072755 and *Staphylococcus haemolytic us* GCA:001611955 only contain one pathogenesis protein (PF04647) that is not present in any *Streptococcus* strain. This protein, PF04647 ArgB, is part of a quorum sensing system. Also *S. saprophyticus* GCA:002209265 only contains one protein not present in any *Streptococcus* strain. This protein, PF05480, is a haemolytic protein unique to *Staphylococcus*. Among *Streptococci, S. parauberis* six *S. iniae* and seven *S. thermophilus* strains lack any pathogenesis protein that separates them from *Staphylococcus*.

4.3.7 Domain shuffling of pathogenic proteins

The *Staphylococcus* pangenome contains 52 domains present in 65 proteins associated to the GO term 'pathogenesis' while the *Streptococcus* pangenome contains 88 domains present in 118 proteins associated to the GO term 'pathogenesis'. 20% of pathogenic proteins in *Staphylococcus* and 25% of the pathogenic proteins in *Streptococcus* consist of a few pathogenesis associated domains combined with domains not directly associated to pathogenesis. This implies that domain shuffling might be an important evolutionary factor for these pathogens. In *Staphylococcus* 46% (30/65) and in *Streptococcus* 72% (85/118) of the pathogenesis associated proteins contain multiple domains. This percentage is much higher than the average percentage of multi-domain proteins of 8.9% and 9.2% for *Staphylococcus* and

Streptococcus respectively. It could be argued that proteins involved in pathogenesis would more often require multiple domains since many of them are cell-wall associated, secreted or contain multiple domains to facilitate interaction between host and the pathogen. The importance of cell wall associated proteins is reflected by the high percentage of 40% of pathogenesis proteins in *Staphylococcus* and 66% *Streptococcus* that contain LPXTG cell-wall anchor domain PF00746. The importance of this domain for pathogenesis was shown in a *S. aureus* mutant with a knockout of *srtA* coding for a class A sortase, which is required for secretion of proteins containing the LPXTG motif. This mutant was unable to form abscess lesions in organ tissues or cause lethal bacteraemia when inoculated in the blood stream of mice [40].

4.3.8 Staphylococcus aureus multi-drug resistance

We investigated the clustering of *S. aureus* genomes in the functional tree associated to the terms "response to drug". We selected the genome of *S. aureus sub species aureus MRSA 252* (GCA:000011505), which is known to be a multiple drug resistant strain [332]. Next, we searched literature for information about drug resistance for eight genomes that cluster together with this strain in the functional tree response to drug. For seven of these strains (JH1, JH9, Mu50, Mu3, T0131, 04-20981), evidence was found for these strains to be multi-drug resistant as well as identifying two pathogenicity islands as the cause of their resistance [333]–[337]. For the last genome (GCA:001640885), no literature or other information could be retrieved. This genome has exactly the same proteins associated to response to drug as the seven strains for which multi drug resistance was reported in literature. Therefore, we can speculate that this strain may have the same multi drug resistance phenotype.

4.3.9 Streptococcus suis pathogenesis zoonotic potential

Large differences in clustering were observed for *S. suis* genomes in the functional trees relating to 'biological adhesion', 'modification of morphology or physiology of other organism' and 'pathogenesis', supplementary material [Additional file 9]. *S. suis* genomes form two groups in the functional tree of biological adhesion, and three groups in the functional tree of pathogenesis and 'modification of morphology or physiology of other organism'.

Similarly, different groups can be distinguished in the PCA plot based on these three functional groups, as shown in **Figure 16 A-C**. We included information from literature on zoonotic species, namely *S. inae, S. agalactiae, S. dysgalactiae S. iniae* and*, S. equi zooepidemicus and S. suis serotype 2 strains* and serotype information and host isolation information for *S. suis* and *S. agalactiae* strains in the labels of **Figure 16 A-C**. Two *S. suis* clusters can be distinguished in the PCA score plot based on proteins related to 'modification of morphology or physiology of other organism' (**Figure 16 B**) and the PCA based on 'pathogenesis' proteins (**Figure 16 C**): the first cluster contains 7 out of the 12 serotype 2 strains, as well as serotype 1, 1,2, 4, 16, while the second cluster contains 5 serotype 2 strains as well as strains with serotype 3, 7, 9 14 and Chz which were all isolated from pigs. The first group contains *S. suis* zoonotic strains of which some are isolated from pig and some from humans. The second group contains are non-zoonotic strains all isolated from pigs.

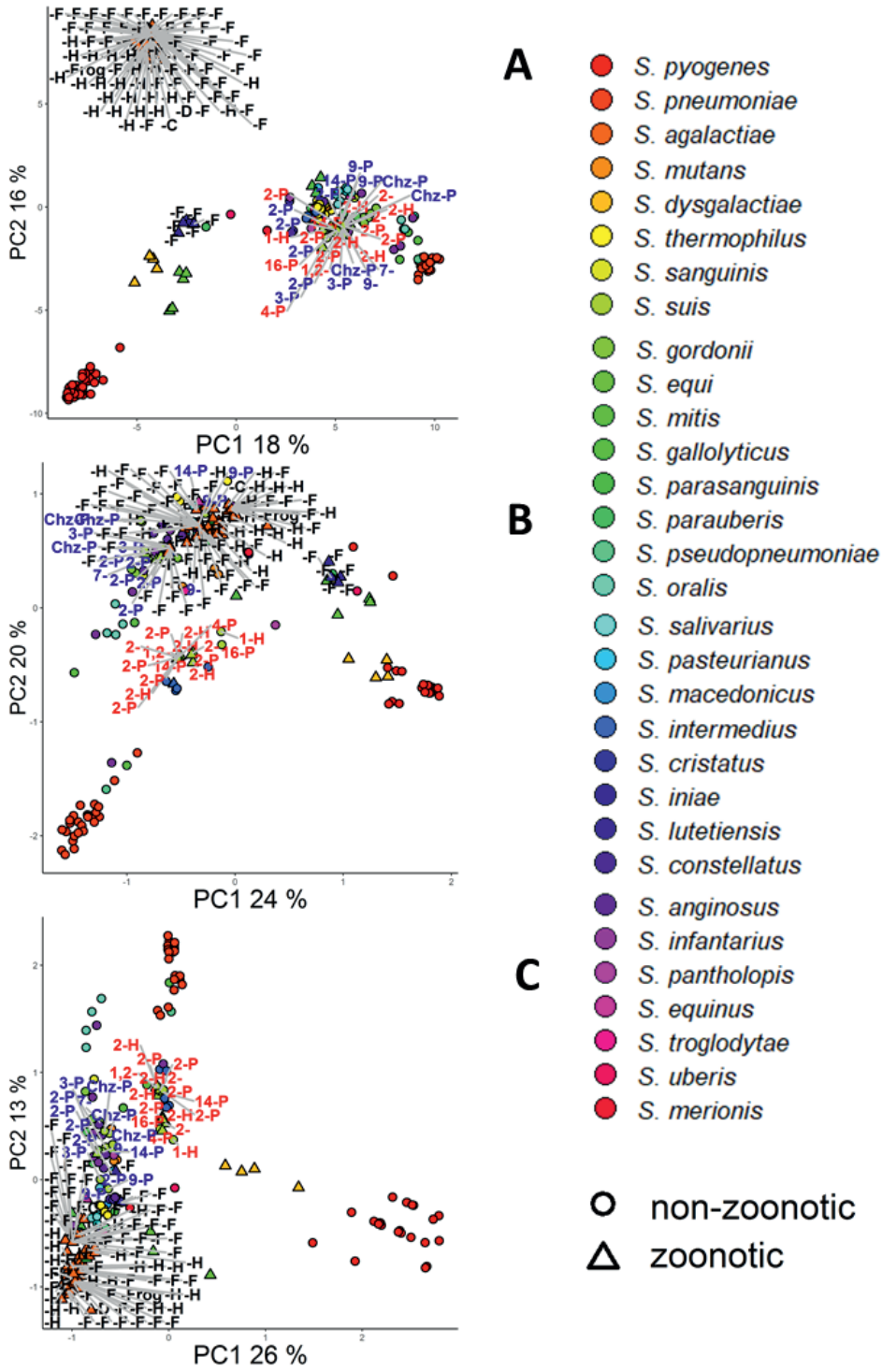

Figure 16. PCA plot of Streptococcus *strains based on all proteins (A), proteins filtered on 'modification of morphology or physiology of other organism' (B) and proteins filtered on 'pathogenesis' (C).* S. suis *serotypes are shown in the label, genomes from species mentioned in literature as having zoonotic capabilities are marked with a triangle and the isolation host is marked in the label with D=dog, F=fish, H=human, P=pig, T=toad. Genomes predicted in this study to have zoonotic potential are coloured red while strains in the cluster predicted not to have zoonotic potential are coloured blue. Fraction of variance explained by each PC is indicated in the axis.*

4.3.10 Streptococcus agalactiae zoonotic potential

Like *S. suis, S. agalactiae* forms two clusters when clustering on GO biological functional groups of proteins. Based on their isolation host, we can see that a cluster contains strains that are zoonotic while the other contains strains that are non-zoonotic. These two groups of *S. agalactiae* strains are better separated when using t-*t*-SNE plots based on all proteins and proteins involved in biological adhesion and pathogenesis (**Figure 17 A-C**) suggesting the existence of few proteins that are present in every genome in each group.

4.3.11 Identification of proteins that confer zoonotic potential

We used Random Forest, a machine learning approach, to investigate the association between genome content and phenotype using 75% of the data for training and 25% of the data for validation. Specifically, presence/absence of proteins filtered on association to GO Biological functions involved in pathogenesis to predict zoonotic potential *S. suis* and *S. agalactiae*, and we investigated which proteins are responsible for the zoonotic potential in these two species. We used functional groups of proteins that were shown to separate zoonotic and non-zoonotic strains for *S. suis* (**Figure 16 B-C**) and for *S. agalactiae* (**Figure 17 B-C**) to train a Random Forest classifier. We investigated their overall importance for prediction as well as their contribution to predicting the class non-zoonotic, and the class zoonotic potential as shown in **Figure 18 A-D**. Where, the 'Impact' measure indicates the relevance of a protein of the prediction of given class. The 'importance' shows the proteins overall importance for the random forest classifier. Random Forest classifiers as well as the optimal hyper parameters can be found in [see Additional file 14].

The protein domain content of the five most important proteins for *S. suis* classification based on 'modification of morphology or physiology of other organism' proteins are: 1) PF01289 a thiol-activated cytolysin, 2) PF17440 thiol-activated cytolysin beta sandwich domain, 3) PF00910 replication initiation protein involved in viral RNA duplication 3) PF00078;PF08388;PF13655 group II intron reverse transcriptase/maturase, 4) PF03432 a relaxase involved in transfer of plasmids, 5) PF00665 Prokaryotic N-terminal methylation motif often found in pilins and other proteins involved in secretion (**Figure 18A**). The most important proteins for *S. suis* classification based on 'pathogenesis' proteins are 1) PF01289 thiol-activated cytolysin, 2) PF17440 a thiol-activated cytolysin beta sandwich domain, 3) PF07564 hypothetical protein containing a domain of unknown function, 4) PF00092; PF00746 chemotaxin protein 5) PF00746;PF08363;PF16364 a glucan binding protein (**Figure 18B**).

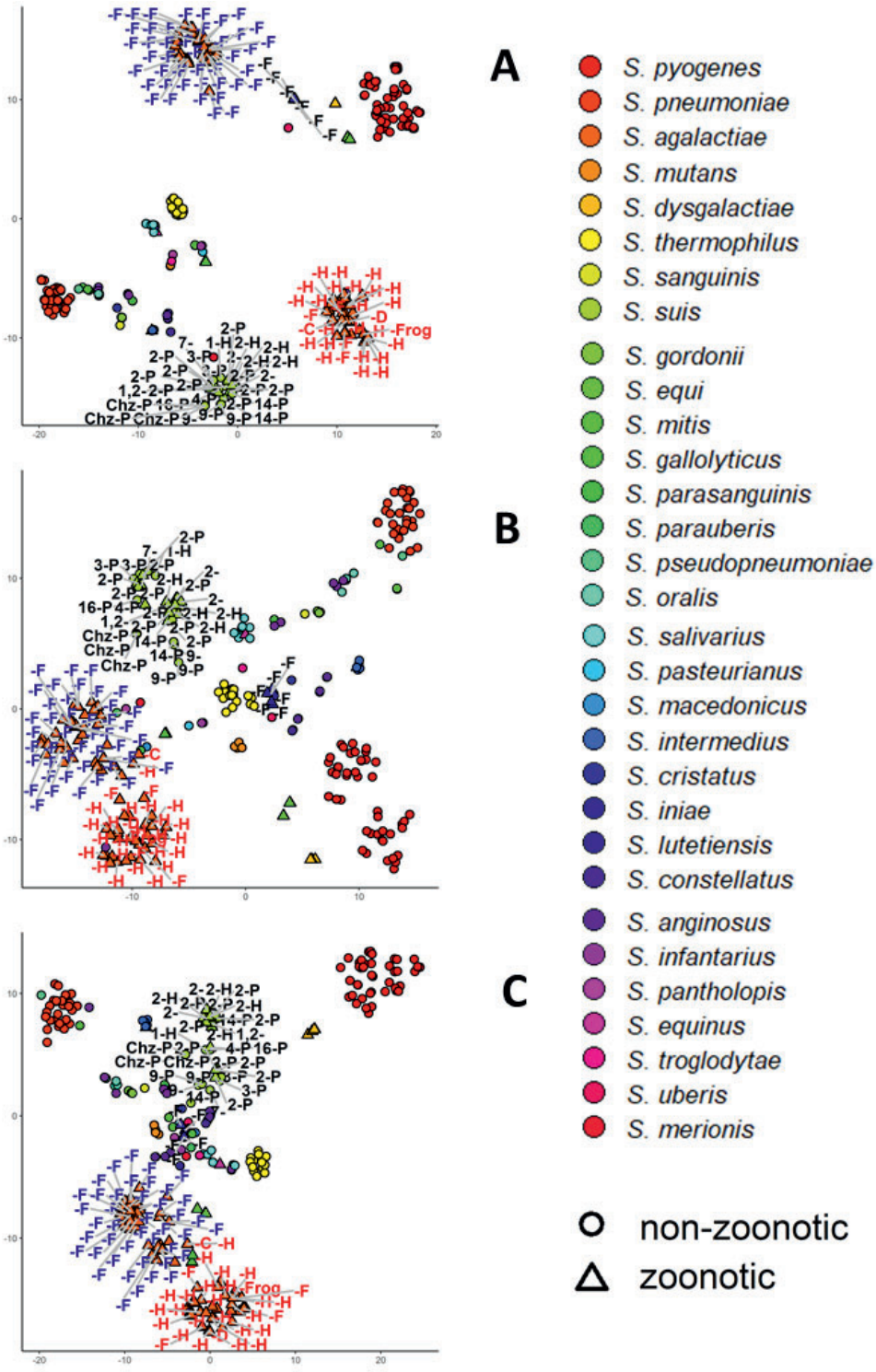

Figure 17. t-SNE plots of Streptococcus *strains based on proteins all proteins (A), proteins filtered on 'biological adhesion' (B) and 'pathogenesis' (C). t-SNE is a technique for dimensional reduction and visualization, so that similar objects appear as nearby objects in the two-dimensional plots here presented.* S. suis *serotypes are shown in the label, genomes from species mentioned in literature as having zoonotic capabilities are marked with a triangle and the isolation host is marked in the label with D=dog, F=fish, H=human, P=pig, T=toad. Genomes predicted in this study to be part of the zoonotic potential cluster are coloured red while strains in the cluster predicted not to have zoonotic potential are coloured blue. Fraction of variance explained by each PC is indicated in the axis*

The *S. suis* classifiers based on 'modification of morphology or physiology of other organism' proteins as well the classifier based on 'pathogenesis' proteins, predict *S. suis* zoonotic potential with 100% accuracy solely based on the presence of either PF01289, a thiol-activated cytolysin or PF17440, a thiol-activated cytolysin beta sandwich **Figure 18 A-B**).

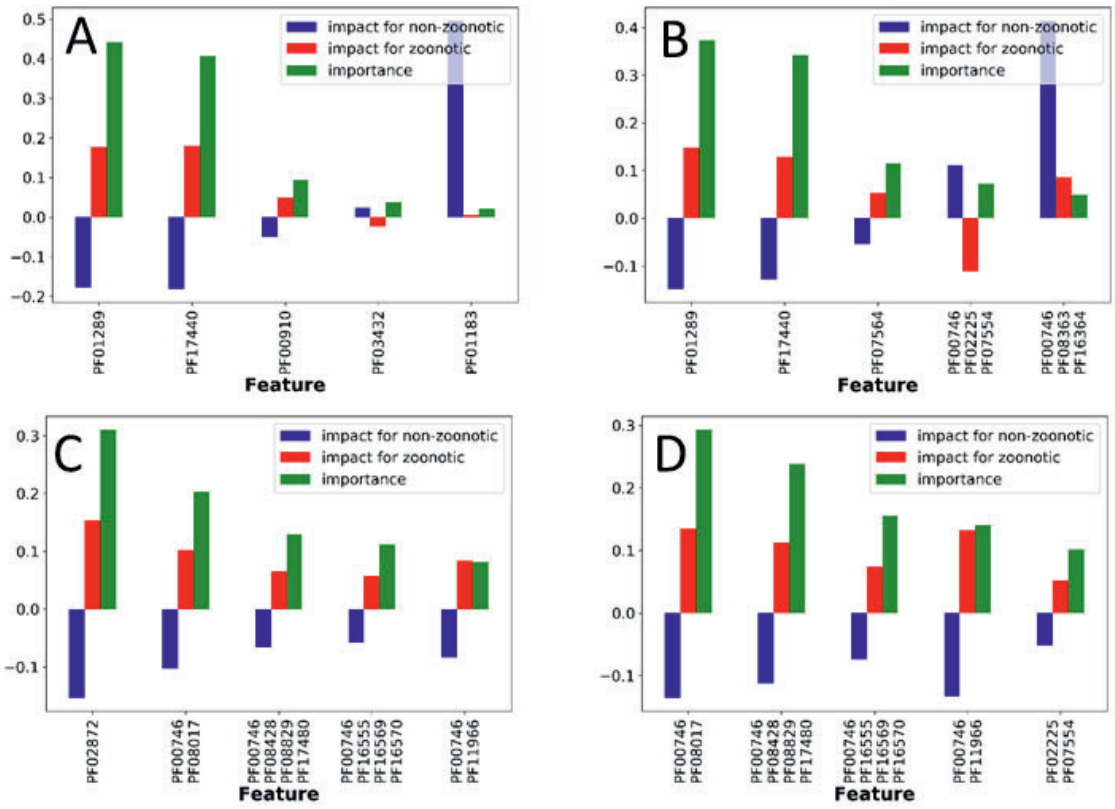

Figure 18. Protein feature contribution to predict the class 'non-zoonotic' and 'zoonotic' as well as the overall importance of the protein feature for classification. A) The five most important 'modification of morphology or physiology of other organism' proteins used to classify S. suis. *B) The five most important 'pathogenesis' proteins used to classify* S. suis. *C) The five most important 'biological adhesion' proteins used to classify* S. agalactiae. *D) The five most important 'pathogenesis' proteins used to classify S. agalactiae.*

The most important features for *S. agalactiae* classification based on 'biological adhesion' proteins are 1) PF02872 a 5'-nucleotidase-C 2) PF00746;PF08017 Fibrinogen binding protein A, 3) PF00746; PF8428; PF08829; PF174802 surface protein Rib and 4) PF00746;PF16555; PF16569; PF16570 pilus complex 5) PF00746;PF11966 a cell wall anchored linked to a ySIRK signal domain (**Figure 18 C**). The most important features for *S. agalactiae* classification based on ´pathogenesis' proteins are 1) PF00746;PF08017 Fibrinogen binding protein , 2) PF00746;PF8428;PF08829;PF17480 2) surface protein Rib and 3) PF00746;PF16555;PF16569;PF16570 pilus complex 4) PF00746;PF11966 a cell wall

anchored linked to a ySIRK signal domain 5) PF02225;PF07554, a serine protease (**Figure 18D**).

For *S. agalactiae* classification, 5'-nucleotidase-C can predict training data with 100% accuracy and test data with 94% accuracy.

4.3.12 Virulence factors of necrotising fasciitis

We compared *Staphylococcus aureus* and *Streptococcus pyogenes* since both are a major cause of (monomicrobial) necrotizing fasciitis. Both *S. aureus* [338]–[340] and *S. pyogenes* [341], [342] fully lyse red blood cells, induce toxic shock syndrome as well as bind and invade epithelial host cells. Based on their GO term association *S. aureus* has 21 proteins associated to pathogenesis that occur in nearly all *S. aureus* genomes and only rarely in any other *Staphylococcus* specie (**Table 10**). Some of these pathogenesis proteins can directly be linked to pathogenesis proteins reported in literature for *S. aureus* [343] and for *S. pyogenes* [344]. An exact match with proteins reported in literature is however not always possible due to differences in annotation. When considering virulence factors that are not unique to *S. aureus* or *S. pyogenes* the number of virulence factors is about 1.5 times as many as reported in literature [345].

Table 10. Proteins associated to S. aureus Pathogenesis (GO: GO:0009405). Domains that are shared between S. aureus and S. pyogenes are underlined. Proteins that are shared have are written in bold.

PROTEIN	DESCRIPTION
PF00746; PF07501; PF17041	LPXTG cell wall anchor; G5 domain, suggested adhesion, in peptide that cleaves IgA; E domain, rod like structure
PF00746; PF17210	LPXTG cell wall anchor, SdrD B-like domain, involved in adhesion to nose squamous cells [346]
PF13545	Crp-like helix-turn-helix domain, possibly cAMP interaction
PF05543	Staphopain peptidase C47, secreted cysteine protease
PF14731	Staphopain proregion
PF03373	Octapeptide repeat, part of SpA virulence factor frequently used to type *S. aureus* strains [347]
PF07968	Haemolysin, part of the Leukocidin/Hemolysin toxin family
PF09199	*Staphylococcal* superantigen-like OB-fold domain, interact with IgA, inhibits the end stage of complement activation and IgA binding to Fc-α-R [348]
PF02216	SpAB protein domain, immunoglobin binding domain
PF02876	*Staphylococcal/Streptococcal* toxin, beta-grasp domain
PF11621	C3 binding domain 4 of IgG-bind protein SB
PF01123	Enterotoxin type B, supertoxin, involved in food poisoning, causing the immune system to release a

	large number of cytokines that lead to significant inflammation
PF03642	MAP domain, major histocompatibility complex class II analog
PF00746;PF01476	LPXTG cell wall anchor motif; LysM domain found in many receptors, peptidoglycan-binding protein [328]
PF00746;PF02986	LPXTG cell wall anchor motif; Fibronectin binding repeat, enables uptake by host cell
PF00746;PF05031	LPXTG cell wall anchor motif; Iron Transport-associated domain, heme and/or hemoprotein-binding
PF07564	Domain of Unknown Function (DUF1542), several proteins containing this domain are involved in antibiotic resistance and/or cell adhesion
PF09023	Staphostatin B inhibits the cysteine protease Staphopain B
PF02821	Streptokinase (SK) is a thrombolytic medication and enzyme, breaks down blood cloths
PF01468	GA module, GA modules may promote bacterial growth and virulence in mammalian hosts by scavenging albumin-bound nutrients and camouflaging the bacteria
PF07554;PF07564;PF08428	FIVAR domain, likely binds fibronectin or more specifically N-acetyl glucosamine, occurs in proteins involved in methicillin resistance; Domain of Unknown Function (DUF1542); Rib/alpha-like repeat. Occurs in some Rib, a thought to confer protective immunity. Occurs in some *Streptococcus* surface proteins. Extracellular matrix-binding protein Ebh

We looked at shared proteins as well as functional alternatives to find the molecular basis for necrotising fasciitis and we used the PFAM description of protein domains as well as description of proteins based on the locus tags associated to these proteins. We found that both *S. aureus and S. pyogenes* contain proteins involved in fibronectin binding, wound invasion, haemolysis, cell adhesion, IgA and IgG binding, multiple (super-)toxins as well as proteins involved in resisting phagocytosis and invading host cells (**Table 10, Figure 15**). For example, PF01123 and PF02876 toxin β-grasp domain together form Enterotoxin type 2 which is important for causing the toxic shock [320], [349]. Enterotoxin type 2 antibodies are currently in clinical trials tested and have shown potential in treating necrotizing fasciitis [350].

S. pyogenes has 10 proteins that present in all *S. pyogenes* species and only occur separately in a few other Streptococcus strains (**Table 11, Figure 15**). Five of these proteins are associated to the ability of *S. pyogenes* to bind to and break down fibrin in blood cloths [342]. Other proteins include toxin and enterotoxin, involved in over-activation of the immune response [351], [352], fibrin binding proteins, involved in adhesion and intracellular access of host cells, as well as proteases, involved in resistance to phagocytosis [353]. *S. pyogenes* has fewer proteins that are unique to this

species compared *to S. aureus* since many other *Streptococcus* strains produce some of the pathogenic proteins present in S. pyogenes [354].

Table 11.

Table 11. Proteins associated to S. pyogenes Pathogenesis (GO: GO:0009405). Domains that are shared between S. aureus and S. pyogenes are underlined. Proteins that are shared are written in bold.

PROTEIN	DESCRIPTION
PF02821	Streptokinase, breaks down blood cloths
PF01640	Peptidase C10 family
PF01123	Enterotoxin type B, super antigen involved in food poisoning
PF02876	Staphylococcal/Streptococcal toxin, beta-grasp domain
PF03734	L,D-transpeptidase catalytic domain, peptidoglycan binding
PF00746; PF01391	LPXTG cell wall anchor motif; Collagen helix, rod like structure, coagulation-fibrinolytic binding in blood, Scl1 adhesin specifically recognizes the wound microenvironment [355]
PF00746; PF02370	LPXTG cell wall anchor motif; M protein repeat, binds IgA, major virulence factor involved in host cell invasion and resistance to phagocytosis [356]
PF00746; PF08017	LPXTG cell wall anchor motif; Fibronogen binding protein, members of this family include the fibrinogen receptor, FbsA which mediates platelet aggregation
PF00746; PF02986	LPXTG cell wall anchor motif; Fibronectin binding repeat, mediate adherence to host cells, enable the colonisation of wound tissue and blood clots
PF00092; PF00746; PF02986	Von Willebrand factor type A domain domains participate in numerous biological events (e.g. cell adhesion, migration, homing, pattern formation, and signal transduction); LPXTG cell wall anchor motif; Fibronectin binding repeat

Only three proteins associated to pathogenesis are shared between *S. aureus* and *S. pyogenes*, Enterotoxin B C-terminal domain (PF02876), Enterotoxin B N-terminal beta-grasp domain (PF01123) and Fibronectin binding protein (PF00746;PF02986). Both fibronectin binding protein A (FnbpA) and B (FnbpB) are expressed during infection conditions and were shown to be complexly regulated by a large number of regulators such as sigma factors and two component systems by Mader et al. [357]. Among these proteins identified in our study are potential biomarkers. FnbpA was found not to be essential in KO studies [358], but was found to be essential in a rapid shotgun antisense RNA method to identify essential genes in *S. aureus*. [359]. No essentiality information is available for FnbpB. S. aureus fibronectin binding protein A (FnbpA) is called fibronectin binding protein X (SfbX) in *S. pyogenes*. For *S. aureus*, FnbpA was found to be essential for entry in the host cells [360]. FnbpA has functional homologs in other species such as *S. epidermidis*, however all homologs lack the C-terminal multiple fibronectin binding repeats variants present in FnbpA, of which at least one high affinity binding repeat is needed for host cell uptake [349], [361], [362]. A SfbX knockout mutants was shown to be only minimally affect *S. pyogenes* ability

to infect epithelial host cells [363]. Enterotoxin type B, is a super toxin involved in over-activation of the immune response and interferes with phagocytosis by supressing the generation of myeloid-derived suppressor cells [364]–[366].

4.4 Discussion

The *Staphylococcus* and *Streptococcus* genera were compared on their genomic properties. Both genera have a similar ratio of their pan and core genome size. It should be considered that this analysis has been done with all fully assembled genomes data that were available at the time of the study. In our study we do not separate between pathogenic and non-pathogenic species since there are several ways to infect humans and animals. Instead, we use the underlying annotation of proteins marked as being involved in GO functions associated to pathogenesis to investigate patterns in pathogenesis. The choice not to define before-hand if species are pathogenic is deliberate since we recognize there are many forms of pathogenesis which depend on both species as well as the infection site as we discuss in the sections "Streptococcus suis pathogenesis zoonotic potential" and in the section "Streptococcus agalactiae zoonotic potential". We do however recognize the selected population affect our results as can be seen for in the ratio of their pan and core genome size found in this study. The alpha value of 1.12 found in our study for *Streptococcus* is higher than the 0.87 values reported by Koehorst et al (12). This difference can be explained by the number of genomes analysed which was 314 in our study opposed to 60 in the study by Koehorst et al. Additionally, we allowed a maximum of one genome per species to be selected in our sampling approach to avoid population bias introduced by species with many sequenced genomes such as *S. aureus* which was not the case in the analysis of by Koehorst et al (12)

Similar to what was found for *Pseudomonas* [325], gene expression variability of essential genes was found to be less than the expression variability of non-essential genes in both *S. aureus* and *S. pyogenes*.

Combination of experimentally determined essentiality and GEM based essentiality prediction were shown to be associated to a higher protein persistence than each of them individually. These results are to be expected since *In vitro* essentiality measurements are often only available for one condition, while GEM can easily be used to predict essentiality over multiple media conditions. Our GEM analysis predicted 153 *Staphylococcus aureus* genes to be essential in 90% of the minimal medium combination tested, while 163 genes were found to be essential for growth on rich medium and minimal medium. For *Streptococcus*, 196 genes were found to be essential in 90% of the minimal medium combinations tested, while no genes were found to be essential on rich medium. The *Streptococcus* model contains exchange reactions for all nutrients necessary for growth, meaning only the biomass reaction was found to be essential. Since we know from experimental results that there are several essential genes in *S. pyogenes*, we chose our method of testing all minimal medium compounds to best balance false positive and false negative results while keeping a unified method for our GEM essentiality analysis in both *S. aureus* and *S. pyogenes*. Similar to what has been experimentally observed and was shown by previous published GEM simulations [324], our simulations show that

Staphylococcus can use amino acids as alternative carbon source for survival in the host [367]. A possible limitation of this approach is that GEM predictions can only be made for metabolic (and their associated) proteins. Although many highly persistent genes tend to be essential, not all are highly persistent. This indicated alternatives in essentiality exist [368]. Similarly, many non-essential genes do have a high persistence, indicating they might be essential for *Staphylococcus* or *Streptococcus* specific functions such as survival and growth in non-lab conditions such as those found in the host.

Differences in pathogenesis, essentiality as well as other properties such as drug resistance, arise from different selection pressures for individual species within genera [369]. For example, similar to what was found in this study, a recent study shows that although streptococcal virulence factors have no clear patterns among species groups, some virulence factors were shown to be congruous with the evolution of species groups [329]. Core genes together with accessory genes form a complex network that comprise the molecular basis of virulence in *Staphylococcus* and *Streptococcus* [370], [371]. Within some individual species, strong selective pressure exist as was shown for *S. aureus* MRSA resistant species where there is an interplay of two strong evolutionary selective pressures: 1) the host type and 2) the antibiotics used in treatment which varies between humans, pets and livestock [372], [373]

We compared the similarity and differences between *Staphylococcus* and *Streptococcus* based on the clustering of species in GO functional trees. Some notable differences where observed between *Staphylococcus* and *Streptoccus*. Many 'modification of host morphology' proteins in *Staphylococcus* are also associated to the GO term 'pathogenesis' while many 'biological adhesion' proteins in *Streptococcus* are associated to the GO term 'pathogenesis'. These results could indicate that modification of host morphology is important for the pathology of *Staphylococcus* strains while biological adhesion is more important for the pathology of *Streptococcus*.

Next, we looked at which pathogenesis proteins separate *Staphylococcus* from *Streptococcus* species. Analysis of the presence and absence of proteins associated to 'pathogenesis' reveals that *Staphylococcus sciuri* GCA:002072755 and *Staphylococcus haemolytic us* GCA:001611955 only contain one pathogenesis protein (PF04647) that is not present in any *Streptococcus* strain. These results could indicate that horizontal gene transfer of pathogenic proteins occurred between *Staphylococcus* and *Streptococcus* or that they only carry pathogenesis proteins derived from a common ancestor.

Additionally, some pathogenic proteins only occur in one or a few genomes, indicating horizontal gene transfer from species outside the *Staphylococcus* and *Streptococcus* genus. Horizontal gene transfer is known to be a driving factor in the development of pathogenesis in *Staphylococcus* and *Streptococcus* [318], [322], [374]–[376]. For example, fibronectin binding domain PF02986 has been acquired by *Staphylococcus* and *Streptococcus* from an animal host, further spread among different *Streptococci* and *Staphylococci* through horizontal gene transfer, and further evolved through domain shuffling [354], [377].

Clustering of *Streptococcus* and *Staphylococcus* species based on different GO functional groups revealed sub cluster to be present for *S. suis* based on GO functional groups *'modification of host morphorlogy'* and *'pathogenesis'* and revealed a sub cluster to be present for *S. agalactiae* based on GO functional groups *'biological adhesion'* and *'pathogenesis'*. What is more, these sub cluster coincides with the potential to infect multiple hosts. It is known that predominantly *S. suis* serotype 2 strains are associated to zoonotic potential [378], [379]. However, as we could see in **Figure 16 A-C**, serotype information is not able to separate zoonotic and non-zoonotic *S. suis* strains.

Based on their isolation host we can see that the first group are *S. suis* strains are zoonotic, while the second group are non-zoonotic strains. Furthermore, human infections with strains for all serotypes in the first cluster have been reported [380], [381]: these results show that these functional groups of proteins can be used to predict the *S. suis* zoonotic potential. Interestingly, the zoonotic group of *S. suis* strains clusters with the human and dog oral commensal *S. intermedius* which can cause meningitis through brain abscesses as well as liver abscesses and in some rare cases endocarditis [382]. Since *S. suis* and *S. intermedius* are distantly related, this clustering is specific for proteins with functions in modification of host morphology and pathogenesis.

The similarity of phenotypes such as causing meningitis and tropism for brain and liver, suggests these traits may be caused by 'modification of morphology or physiology of other organism' and 'pathogenesis' proteins, and suggest a causal relationship between the proteins associated to these GO terms and the observed phenotype.

Investigation of proteins required to predict *S. suis* zoonotic potential using a random forest classifier revealed PF01289 thiol-activated cytolysin, 2) PF17440 a thiol-activated cytolysin beta sandwich domain to be the two most important factors associated to zoonotic potential. In support of these findings, it was found that a *S. suis* cytolysin knockout mutant made the strain non-haemolytic and non-cytotoxic for cultured macrophage-like cells [383] while increased secretion of thiol activated cytolysins was shown to directly cause epithelial cell damage in humans, allowing *S. suis* to spread into deeper tissues [384]. Based on these studies it appears these two cytolysins are involved information of a pore-forming complex in cholesterol containing host membranes, which explains their importance for conferring zoonotic potential.

It has been suggested that *S. agalactiae* may have jumped from animals to humans in a certain moment of the evolution although it is still debatable if this zoonotic potential remains nowadays [15]. Here, we show that based on their genomic content *S. agalactiae* can be separated in two groups, one that is zoonotic and infects humans, fish, and dog, and one group that only infects fish. This separation can be made based on all proteins, indicating zoonotic and non-zoonotic species are likely to have separated some time ago. In the *t*-SNE plots of biological adhesion and pathogenesis a third group can be seen of strains that infects mainly fish but also cow and human. This third cluster contains *S. agalactiae* strains that infect human most likely originate from this cluster and have further adapted to their human host by acquisition of proteins involved in biological adhesion and pathogenesis. The strains in this group

appear to retain zoonotic potential since the cluster contains isolates from humans, fish, and dog.

Investigation of proteins required to predict *S. agalactiae* zoonotic potential using a random forest classifier revealed multiple proteins to be important for classifying species as zoonotic. *S. agalactiae* 5'-nucleotidase-C is present in two proteins: Trifunctional nucleotide phospho-esterase protein YfkN precursor and Endonuclease YhcR precursor. Secreted nucleases play a role in evasion of the human innate immune response via destruction of extracellular traps and interference with phagocytosis signals [385]. Fibrinogen binding protein A allows *S. agalactiae* to attach to fibrinogen and to aggregate platelets [386]. Rib protein contains a Rib domain that confers protective immunity and an alpha C and alpha N protein domains involved in invasion and translocation along human epithelial cells according to their PFAM description. The pilus complex contains Pillin D1, Pillin B and Pillin D3 domains and contributes to the initial attachment and invasion of lung and cervical epithelial cells [386]. PF02225;PF07554 CspA serine protease breaks down three chemokines that attract and activate neutrophils [387].

In summary, all proteins important for classification of zoonotic potential appear to be causal to the zoonotic potential phenotype. Of these proteins, nucleosidases YfkN, YhcR and fibrinogen binding protein appear to be the most important factors for *S. agalactiae* zoonotic potential.

4.5 Conclusions

In this study we dissected *Staphylococcus* and *Streptococcus* pathogenesis through the systematic and integrated analysis of genomic, functional, metabolic, and expression data. Both genera were found to have a closed pangenome and lower expression variation for essential and highly persistent genes than for non-essential and low persistent genes. The study of functional groups of proteins in the pangenome of *Staphylococcus* and *Streptococcus* involved in pathogenesis, indicates that domain shuffling and horizontal gene transfer have played an important role in the development and acquisition of pathogenesis proteins of *Staphylococcus* and *Streptococcus* species.

The analysis of bacterial clusters based on functional groups of proteins involved in pathogenesis shows that clustering of strains correlates with phenotypes such as zoonotic potential. Comparison between *S. aureus* and *S. pyogenes* indicate three proteins, Enterotoxin B C-terminal domain, Enterotoxin B N-terminal beta-grasp domain together with several functionally equivalent proteins allow *Staphylococcus aureus* and *Streptococcus pyogenes* to cause necrotizing fasciitis.

We have also shown that prediction of the phenotype zoonotic potential only requires information about a few proteins, suggesting a direct causal relationship with zoonotic potential. These findings will enable further research in each of the areas addressed, whereas the approaches and methods herein deployed provide a solid basis towards large-scale prediction of phenotypes based on genomic information.

4.6 Methods

4.6.1 Genome retrieval and annotation

All available completely assembled genomes of 235 *Staphylococcus* and 315 *Streptococcus* strains were downloaded as EMBL files from EBI-ENA using the Python EnaBrowserTool [388]. Lists of these genomes accession number, name and taxon ID can be found in supplementary material [Additional file 1&2]. Genome EMBL files were converted to RDF and *de novo* annotation was performed storing the results in a graph file per genome using SAPP, a Semantic Annotation Platform with Provenance [65] and the GBOL ontology [389]. Gene calling was performed using Prodigal with codon table 11 [390]. Annotation was performed using InterProScan version 5.25 [391]. Protein domains were identified by InterProScan by their Pfam identifier [42]. The GNU "parallel" package was used to perform all of the above steps in parallel [65].

The graph files were loaded in GraphDB Free version 8.4.1 in order to query the annotated genomes. Additionally, taxonomic information from UniProt was downloaded in RDF format and loaded in GraphDB. The GraphDB SPARQL endpoint was queried using the Python SPARQLWrapper [392] package and the R Curl package [393] to retrieve information and store them as matrixes given in supplementary material [see Additional file 3&4]. These files were used for all subsequent analyses.

4.6.2 Estimation of the size of the pan- and core genome

Proteins were compared based on their Pfam domain content. We defined protein domain content as the alphabetical order of all unique domains associated with a given protein. A matrix was built to collect information on the presence or absence of proteins in each genome. Two sampling approaches were used: 1) genomes were randomly selected from all genomes in the analysed genera and 2) a maximum of one genome per species was selected to avoid bias introduced by species with many sequenced genomes. One up to the total number of genomes were sampled and analysed using the micropan R package [394] to investigate the effect of the number of genomes on the estimation of the size of the pan- and core genome. Additionally, these samples we used to estimate the sizes of the pan- and core genome using a binomial mixture model using the micropan BinomixEstimate function with 5000 permutations and a core detect probability of 1. The process was repeated 10 times to estimate the variance of the estimated size of the pan- and core genomes. The Heaps' function was used to fit a Heaps' regression model; $\alpha > 1$ indicates convergence of the size of the pan-genome and that it is closed.

4.6.3 Variability of gene expression and its association to persistence

Gene variability was calculated based on 156 *S. aureus* RNA samples from 44 conditions ranging from laboratory to conditions mimicking infection, measured by Tiling arrays [357]. These 44 conditions can be categorized in four groups: 1) rich medium (TSB), 2) minimal medium (CDM), 3) cell culture media (RPMI, pMEM) 4) in human plasma (plasma), 5) growth with human bronchial epithelial cell line S9 and the human monocyte cell line THP-1.

Samples were taken at different time points and for infection simulations oxygen availability was limited at later time points. For a complete description of the conditions we refer to S1 Data in the original paper by U. Mader et al [357]. For every gene we considered its expression profile over all samples and a variability value was calculated as the ratio between the standard deviation and the mean expression value using the same approach as in Koehorst *et al.* [325].

4.6.4 Protein persistence and essentiality

We defined the persistence of a gene as

$$Persistence = \frac{N\,(orth)}{N}$$

where *N (orth)* is the number of genomes carrying a given orthologue and *N* is the number of genomes searched [395]. Orthologue genes were identified as genes with identical protein domain content. Locus tags associated to the genes were inferred from the original annotation and used to integrated genome wide gene essentiality data from transposon mutagenesis studies for *Staphylococcus* strains S0385 grown on whole porcine blood [396], NCTC8325 Newman grown on BHI broth [358] and JE2 grown on Handke mannitol medium [397] and *Streptococcus* strains *S. pyogenes* M1T1 strain 5448 and M49 strain NZ131 grown in rich Todd-Hewitt Yeast (THY) medium [398].

4.6.5 GEM-based predictions of essentiality

Gene essentiality analysis based on genome scale modelling was performed using the genome-scale, constraint-based metabolic model (GEM) of *S. aureus* NTCTC 8325 [324] and the GEM model of *S. pyogenes* M49 [399]. First, a minimal medium was determined using the 'cobrapy minimal_medium function'. All carbon, nitrogen, sulphur and phosphor sources from the medium that could support growth were detected by substituting the default carbon, nitrogen, sulphur and phosphor sources. All combinations of minimal media containing these carbon, nitrogen, sulphur and phosphor sources were generated.

Gene essentiality for all combinations of minimal media containing these carbon, nitrogen, sulphur and phosphor were tested by performing single gene deletions followed by flux balance analysis optimizing for growth. If a gene knock-out reduced predicted growth for the media compositions below 1% the gene was considered conditionally essential. Genes predicted to be conditionally essential in at least 90% of the in-silico media compositions were marked as essential. All optimizations were performed using the Gurobi optimizer 8.1 [400] with COBRApy 0.13.4[401] and Python 3.6.

4.6.6 Functional Analysis

Genome information was retrieved from associated literature and from the Biosample database [402], including serotype information and zoonotic potential and isolation-host. Zoonotic classification was derived from literature for *S. inae* [403]–[406], *S.*

agalactiae [15]*, S. dysgalactiae, S. equi* [407]–[409] *and at the serotype level for S. suis* [410]–[412]. For all zoonotic *Streptococcus* species data about the isolation host was retrieved from the Biosamples databases [413] or literature [379], [380], [421], [381], [414]–[420]. Additional Gene Ontology (GO) annotation from the GODM (GO Domain Miner) database [345] was added to proteins based on their domain content, increasing the number of GO terms by approximately 10-fold compared to GO term annotation retrieved from the InterPro database. Literature was used to select 17 GO terms in the Biological process ontology with known or suspected association to pathogenesis [422]–[426] (**Error! Reference source not found.**).

The presence/absence matrix of proteins was filtered on proteins annotated with any of the 17 GO terms (**Error! Reference source not found.**) or their descendent GO terms using the R GO.db package [427]. The filtered matrix was used to calculate the Euclidean distance between genomes. Hierarchical complete-linkage clustering was used to generate dendrograms. These GO-specific dendrograms were compared to a reference dendrogram based on all proteins.

Because these dendrograms are based on annotation of proteins for a specific function, we will refer to them as 'functional trees'. Euclidean distances of genomes in the functional trees and the reference tree were calculated and scaled to values between 0 and 1 using the R scale function using the minimum value for centring, and (min – max) for scaling. Scaled values were used to calculate similarity scores for the position of genomes in each functional tree compared to the reference. These similarity scores were calculated as the Pearson correlation between the scaled Euclidean distances of genomes in the functional tree and the scaled Euclidean distance in the reference tree. Interactive heatmap were generated showing the presence and absence of proteins per genome, while showing the similarity in the side column to highlight differences compared to the reference tree. These interactive graphs were generated using the dendextend and heatmaply packages [428], [429]. Similarity scores for functional trees were calculated using the dendextend cor_cophenetic function.

Matrix manipulations, Principal Component Analysis (PCA), t-Distributed Stochastic Neighbour Embedding (t-SNE) and graphs were performed using R 3.6.1 [430], the prcomp command, and the Rtsne 0.15 [431] and ggplot2 2.3.2.1 [432] packages. t-SNE was performed with default parameters.

4.6.7 Random Forest classification

Proteins belonging to GO categories 'pathogenesis', 'modification of morphology or physiology of other organisms' and 'biological adhesions', were used to train Random Forests classifiers for *S. suis* and *S. agalactiae* strains to predict whether they belong to the class 'zoonotic potential' on 'non-zoonotic potential'. This classification was based on the clustering of in PCA and t-SNE plots which revealed the presence of a zoonotic and a non-zoonotic group of strains. Data was split in 75% training data and 25% validation data.

Data was loaded using Python 3.6, pandas 0.24.2. Skicit-learn 0.20.3 used to load data and train Random Forest classifiers. Treeinterpreter 0.1.0 was used to interpret

feature (protein) importance for classification in general and feature contribution to predict specific classes. Grid search for 300 combinations of parameters was performed optimizing the parameters n_estimators, max_features, max_depth, min_samples_split, and the min_samples. Iterative rounds of feature reduction, that is removal of the protein which least contribute to the classification, followed by hyper parameter optimization, was used to find the minimal set of features (proteins) needed to classify both training and test data. Feature importance and contribution were plotted using matplotlib 3.0.3.

4.7 Supplementary material

All supplementary information is available at:

https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-021-07388-6#Sec25

Additional file 1 - List of Staphylococcus genomes (XLSX 15 KB)

Additional file 2 - List of Streptococcus genomes (XLSX 18 KB)

Additional file 3 - Staphylococcus All proteins (TSV 1.3 MB)

Additional file 4 - Streptococcus All proteins (TSV 1.9 MB)

Additional file 5 - Estimated Pan and Core genome size, Heaps Analysis (PDF 367 KB)

Additional file 6 - Staphylococcus Functional trees (PDF 495 KB)

Additional file 7 - Staphylococcus PCA (PDF 7.3 MB)

Additional file 8 - Staphylococcus t-SNE (PDF 6.4 MB)

Additional file 9 - Streptococcus Functional trees (PDF 561 KB)

Additional file 10 – Streptococcus PCA (PDF 6.6 MB)

Additional file 11 - Streptococcus t-SNE (PDF 8.3 MB)

Additional file 12 – Staphylococcus & Streptococcus combined PCA (PDF 6.4 MB)v

Additional file 13 - Staphylococcus & Streptococcus combined t-SNE (PDF 5.9 MB)

Additional file 14 Optimal hyper parameters (TXT 1 KB)

All Supplementary files, Figures as well as additional code not present in the supplementary files of the published manuscript are available at: https://github.com/NielsZondervan/PhD Thesis.

Acknowledgements

We would like to thank Jianan Chen for performing the initial data exploration for the genome comparison of *Staphylococcus* and *Streptococcus*.

4.8 Authors' contributions

NZ performed the main analyses and wrote the draft manuscript. MSD and ES participated in the design of the study and supervised and directed the research. MSD and ES and VdMS revised the manuscript. All authors contributed to the writing of the final version of the manuscript. The authors read and approved the final manuscript.

4.9 Funding

This work has been supported by European Union through the FP7 programme under grant agreement No. 305340 (INFECT), the SysteMTb project (HEALTH-F4–2010-241,587) and the Horizon 2020 research and innovation programme under grant agreement No. 634942 (MycoSynVac) and from The Netherlands Organisation for Health Research and Development (ZonMW) through the PERMIT project (Personalized Medicine in Infections: from Systems Biomedicine and Immunometabolism to Precision Diagnosis and Stratification Permitting Individualized Therapies, project number 456008002) under the PerMed Joint Transnational call JTC 2018 (Research projects on personalised medicine - smart combination of pre-clinical and clinical research with data and ICT solutions). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Chapter 5
Predicting *Mycoplasma* tissue and host specificity from genome sequences

Submitted for publication:

Niels A. Zondervan, Vitor A. P. Martins dos Santos, Maria Suarez-Diez. "Predicting *Mycoplasma* tissue and host specificity from genome sequences". Preprint available online at https://biorxiv.org/cgi/content/short/2022.08.08.503189v1.

5.1 Abstract

This paper investigates which Mycoplasma proteins are most predictive of tissue and host trophism and to which functional groups of proteins they belong. We retrieved and annotated 432 *Mycoplasma* genomes and combined their genome information with host and tissue isolation data. We compared clustering of *Mycoplasma* and *M. pneumoniae* strains based on different functional groups of proteins. We found that proteins belonging to the Gene Ontology (GO) Biological process group '*Interspecies interaction between organisms*' proteins are most important for predicting the pathogenesis of *Mycoplasma* strains, while those belonging to *'Quorum sensing'* and *'Biofilm formation'* proteins are most important for predicting pathogenesis of *M. pneumoniae*.

Two Random Forest Classifiers were trained to accurately predicts host and tissue specificity based on only 12 proteins. For *Mycoplasma* host specificity CTP synthase complex, magnesium transporter MgtE, and glycine cleavage system are most important for correctly classifying *Mycoplasma* strains that infect humans including opportunistic zoonotic strains. For tissue specificity, we found that a) known virulence and adhesions factor Methionine sulphate reductase MetA is predictive of urinary tract *infecting Mycoplasmas* b) an extra cytoplasmic thiamine binding lipoprotein is most predictive of gastro-intestinal infecting *Mycoplasmas* and c) a type I restriction endonuclease is most predictive of respiratory infecting *Mycoplasmas* and d) a branched-chain amino acid transport system is most predictive for blood infecting *Mycoplasmas*.

Keywords

Mycoplasma; trophism; zoonotic; zoonosis pathogenic; traits; phenotype; prediction; machine-learning; random-forest; classifier

5.2 Introduction

Mycoplasmas are bacteria that adapted to live in their host environment through genome reduction, resulting in a small genome and small cell size [433], [434]. Despite their small genome size, *Mycoplasmas* still contain many genes that are not essential but enhance growth in the various host conditions encountered [368], [435], [436]. *Mycoplasmas* have greatly reduced metabolic capabilities growing only in the fastidious conditions of their selective host [20]. This makes it hard to grow *Mycoplasma* on serum free defined media. For *M. pneumoniae,* a defined media was developed by analysing their membrane components and metabolic capabilities and adding those lipids to the medium that normally are directly recruited from the host environment [21], [437], [438].

Because of their strong host adaptation, pathogenesis of *Mycoplasmas* is hard to typify since their ability to infect and survive in a host is largely a systematic property of their obligatory pathogenic lifestyle, and not the result of a well-defined set of virulence proteins. Only few *Mycoplasma* proteins are directly categorised as pathogenic based on their GO Biological Function annotation. Examples of these proteins are *M. pnueumoniae* CARD toxins, adhesins, motility proteins and hydrogen peroxide production which are directly associated to virulence [439], [440]. These virulence proteins are however not essential, while many metabolic proteins such as glycerol metabolism proteins GlpF *and* GlpK are essential for *M. pneumoniae* growth in host conditions [441].

Previously, clustering based on functional groups of proteins was successfully used to predict pathogenic traits such as zoonotic potential for *Streptococcus suis* and *Streptococcus agalactiae* [44]. Here we use a similar approach combined which systematic collection of meta-data from the BioSamples database [442], to identify those proteins predictive of *Mycoplasma* host and tissue infection types. The BioSample database contains tissue and host isolation data for a larger number of *Mycoplasmas* than was previously available for *Streptococcus suis* and *Streptococcus agalactiae,* allowing for better training and validation of machine learning models. We adjusted our approach to *Mycoplasmas* by selecting those functional groups of proteins known from literature to be important for pathogenesis of *Mycoplasmas*. We compared clustering of 430 *Mycoplasmas* based on 19 Gene ontology (GO) Biological process categories of proteins associated with pathogenesis to identify functional groups of proteins important for *Mycoplasmas* in general as well as *M. pneumonia* pathogenesis. We combined this approach with random forest classification to accurately predicts 3 host and 4 tissue isolation sites for *Mycoplasma* genomes and to identify those proteins important for each host and tissue type.

5.3 Materials and methods

5.3.1 Genome retrieval and annotation

We retrieved 430 completely assembled *Mycoplasma* genomes from EBI-ENA using the Python EnaBrowserTool [388]. A list of these genomes can be found in Supplementary material 1A. Semantic Annotation Platform with Provenance (SAPP) [65] and Genome Biology Ontology Language (GBOL) [389] were used to perform *de novo* annotation and to store annotated genomes as graph files. Gene calling was performed using Prodigal 2.6.3 with codon table 4 [390]. Protein domains were identified by InterProScan 83.0 by their Pfam identifier [42]. The GNU "parallel" package version 20161222 was used to perform all the above steps in parallel [65]. Graph files were loaded in GraphDB Free version 9.7.0. Additionally, taxonomic information from UniProt was downloaded in RDF format and loaded in GraphDB. The GraphDB SPARQL endpoint was queried using the Python SPARQLWrapper [392] package and the R Curl package [393]. Genes are annotated by their protein signature, which we defined as the protein PFAM domains present in the protein. We defined such signature by ascendingly ordering all domains present in a protein concatenated with a ";" between the different domains.

5.3.2 Functional Analysis

Meta data for 430 genomes and 187 species was retrieved from the Biosamples [442] database using their API and was stored in GraphDB. The Biosample data was normalised by combining different metadata fields and standardizing the labels used for host and tissue types and was combined with taxonomic information from UniProt. Gene Ontology (GO) annotation from the GODM (GO Domain Miner) database [345] was added to protein annotation based on their domain content. The GraphDB SPARQL endpoint was queried using the Python SPARQLWrapper [392] package and the R Curl package [393] to retrieve information and store them as tab-separated files. These tab-separated files were used for all subsequent analyses. We used a literature study to identify 19 GO Biological process ontology terms with known or suspected association to pathogenesis [422]–[426].

M. pneumoniae genomes were overrepresented in the dataset (165 out of 430 genomes), therefore we reduced the number of *M. pneumoniae* genomes to 12 randomly selected genomes. Functional trees were build based on each of the 19 GO Biological functional groups of proteins as well as a reference tree based on all proteins. We analysed the similarity in cophenetic clustering of these 19 functional trees using the R dendextend package version 1.13.2. In addition, we repeated the clustering of these functional trees using only the 165 *Mycoplasma pneumoniae* genomes. The results of the cophenetic clustering of these functional trees based on all *Mycoplasmas* and *M. pneumoniae* were compared and plotted using the R pheatmap package version 1.0.12.

5.3.3 Random Forest classification

Two Random Forest classifiers were trained using sklearn version 0.24.2 to predict host and tissue trophism of *Mycoplasma* species. The host classifier was trained to predict the three host classes: *human, pig-boar, ruminant*. The tissue isolation site classifier was trained to predict 4 isolation sites: *respiratory, blood*, *gastro-intestinal track*, and *uri-genital track*. The data was cleaned by removing all classes with less than 5 instances before splitting the data into 75% Training and 25% Test samples. The classifier was scored based on *'f1_macro'*, meaning that for each class, the performance is weighted equally irrespective of the number of samples, scoring the classifier while balancing false positives and negatives.

Hyper parameter optimization was performed using a grid search for 300 combinations of the parameters *n_estimators*, *max_features*, *max_depth*, *min_samples_split*, and *min_samples*. Ten times Cross validation and out of bag samples were used to avoid overfitting on the training data. The trained Random Forest classifiers were used as basis to iteratively reduce features, using Treeinterpreter 0.1.0 to interpret feature (protein) importance for overall classification as well as for specific tissue and host classes. For each iteration, the least important feature was removed until a set of the 12 most predictive features (proteins) were left. Lastly, we performed a second round of grid hyper parameter optimization when using the reduced set of protein feature. The heatmaply R package version 1.0.12 was used to plot feature importance and contribution.

5.4 Results & Discussion

The 436 annotated genomes contained 2306 unique proteins, of which 10% were multi-domain proteins. Genomic information was merged with GO Biological function annotation for proteins and with sample isolation metadata for the 436 *Mycoplasma* genomes from the BioSamples database. Host isolation data was available for 394 genomes and tissue-isolation information was available for 157 genomes. We normalized the metadata by combining different metadata fields standardizing labels to describe the same infection host and tissue in different samples. *M. pneumoniae* was overrepresented in the original data set with 165 out the 430 genomes. Therefore, we limited the number of *M. pneumoniae* genomes to 12 randomly selected genomes in our *Mycoplasma* dataset resulting in a dataset of 277 genomes. The 165 *M. pneumoniae* genomes were kept as a separate dataset.

5.4.1 Clustering based on GO functional groups of proteins

Based on literature research, we created a list of 19 GO categories expected to be associated to pathogenesis. We investigated the cophenetic distance between phylogenetic trees based on these 19 GO categories as well as a reference tree based on all proteins. The cophenetic distance is a measure of distance between genomes that have been clustered in two dendrograms providing a single similarity score between two trees. By combining all pairwise cophenetic distance scores we can create a heatmap showing the (dis)-similarity of trees based on the 19 GO functional groups of proteins (see **Figure 19**). A list of the 19 GO IDs with their labels can be found in **Table 12**.

We compared the heatmap of the similarity of these 19 trees based on the 277 *Mycoplasma* genomes with the similarity of the 19 trees based on 165 *M. pneumoniae* genomes. The objective of this comparison was to learn which GO functional groups of proteins have the highest similarity in clustering of genomes based on *'Pathogenesis'* proteins cluster. By performing the clustering both for *Mycoplasma* genomes in general as well as for *M. pneumoniae* genomes only, we identified which functional groups of proteins are likely to be important for *Mycoplasma* pathogenesis as well as which functional groups of proteins are most likely to be important for *M. pneumoniae* pathogenesis (**Figure 19**).

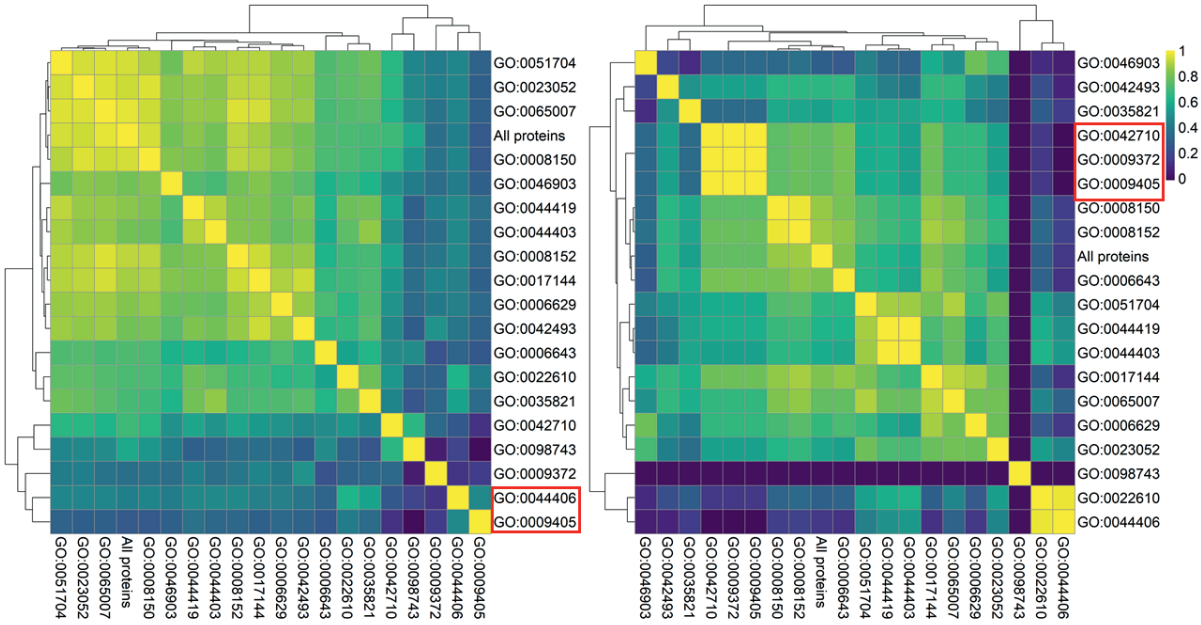

Figure 19. Left cophenetic correlation of GO functional trees based on All Mycoplasma. Right Cophenetic clustering based on M. pneumoniae genomes only

Table 12. GO ID's associated to Pathogenesis together with their names.

GO ID	DESCRIPTION
GO:0008150	Biological process
GO:0008152	*Metabolic process
GO:0006629	Lipid metabolic process
GO:0006643	membrane lipid metabolic process
GO:0017144	Drug metabolic process
GO:0042493	Response to drug
GO:0023052	*Signalling
GO:0065007	*Biological regulation
GO:0022610	*Biological adhesion
GO:0044406	Adhesion of symbiont to host
GO:0051704	Multi-organism process
GO:0044419	Inter species interaction between organisms
GO:0042710	Biofilm formation
GO:0098743	Cell aggregation
GO:0044403	Symbiont process
GO:0009372	Quorum sensing
GO:0035821	Modification of morphology or physiology of other organism
GO:0009405	Pathogenesis
GO:0046903	Secretion

As can be seen in **Figure 19**, for *Mycoplasma,* the tree based on proteins belonging to the GO category '*Inter species interaction between organisms*' (GO:0044419) has

the highest similarity with the tree based on proteins annotated with the GO term *'Pathogenesis'* (GO:0009405). We postulate that this high similarity might indicate that pathogenesis of *M. pneumoniae* involves proteins with functions in '*Symbiont or Multi interspecies interactions'*. Further investigation shows that this similarity was due to a high overlap in annotation, resulting in the same proteins being in both functional groups of proteins. All 26 *Mycoplasma* proteins annotated with the GO term *'Pathogenesis'* contained at least one domain, which is also present in the 147 proteins annotated belonging to '*Interspecies interaction between organisms'*. This overlap in annotation shows that these 26 proteins from the GO Biological functional group '*Interspecies interaction between organisms*' is at least important for *Mycoplasma* pathogenesis.

For *M. pneumoniae,* the trees based on '*Quorum sensing'* (GO:0009372) and *'Biofilm formation'* (GO:0042710) on proteins belonging to the GO category proteins have the highest similarity (1.0) as the tree based on proteins belonging to the GO category '*Pathogenesis*'. Not a single protein in these three categories is present in any of the other three GO categories, ruling out overlap in annotation as the reason for the high similarity of the trees. The high similarity in clustering might therefore indicate that proteins in the categories '*Quorum sensing'* and *'Biofilm formation'* are important for *M. pneumoniae's* pathogenesis.

Multiple studies support the notion that biofilm formation is important for *M. pneumoniae*'s pathogenesis. For example, Type 1 and Type 2 *M. pneumoniae* which have different phenotypes have different biofilms [443]. Biofilm formation is implicated in in chronic infections, with *M. pneumoniae* cells aggregation being important for infections [444]. We found no studies confirming the importance of '*Quorum sensing'* sensing for *M. pneumoniae* pathogenesis. However, it would not come as a surprise if Quorum sensing would be important for *M. pneumoniae* pathogenesis since quorum sensing plays an important role in pathogenesis of other lung infecting bacteria such as *Streptococcus pneumoniae* [445] and *Klebsiella pneumoniae* by regulating virulence systems such as ESX-3, biofilm formation, and secretion of PgaA porin [446], [447]. *M. pneumoniae* virulence systems such as CARD toxins are upregulated when in contact with host cells and in acidic conditions. We postulate that quorum sensing proteins are likely to be involved in pathogenesis by for example sensing host conditions, cell to cell contact to regulate motility, and virulence.

5.4.2 Zoonotic potential

Only two zoonotic strains were identified in our dataset, GCA:001005165 and GCA:000012765, belonging to the *M. capricolum* species group. Two out of the 13 strains were isolated from humans, 2 from goats and 1 from a Tibetan Antilope. The remaining 8 *M. capricolum* genomes have no host isolation data available. We see that *M. capricolum* is taxonomically mixed with *M. leachii* which infects cow. The other M. capricolum isolates are mostly from the respiratory tract, while one zoonotic strain was found in the bloodstream. For the other, no tissue isolation data is available. We hypothesize that *M. capricolum* zoonotic capability is likely limited to infecting the blood stream. In general, it appears that *Mycoplasma* species are so adjusted to their host that they have a limited zoonotic potential [433].

5.5 Predicting host and tissue trophism

Two random forest classifiers were built to predict *Mycoplasma* strains host and tissue infection site respectively. The data was filtered on host classes with a minimum of 5 strains associated to them, resulting in 125 genomes and 3 classes for the host classifier: *human, pig-boar, ruminant*.

Similarly, data was filtered on tissue classes with a minimum of 5 strains associated to them, resulting in 91 genomes and 4 classes for the tissue classifier: *blood, gastro-intestinal, respiratory, and uri-genital tracks*.

The resulting datasets were separated in 75% training and 25% test data. Models were fitted with all protein features using hyper parameter tuning followed by iterative feature reduction to select a set of the 12 most important protein features for classification. A final round of hyper parameter optimization was performed. Models were trained giving equal weight to each class to consider unequal numbers of the classes in both training and test data.

The resulting classifiers predict host and tissue specificity with a high precision on independent dataset not used to train the classifiers (**Table 13**). An overview of optimal hyper parameters as well as the scores for the classifier for both the full and the reduced set of features 12 features, and confusion matrixes can be found in Supplementary file 1B.

Table 13. Classifier score on independent test data using the 12 most important protein features.

	SCORE HOST CLASSIFIER	SCORE TISSUE CLASSIFIER
PRECISION	0.94	0.89
RECALL	0.89	0.88
FSCORE	0.91	0.88

The classifiers predict the provided classes with great accuracy, precision, and recall using only 12 features. The 12 protein features were analysed for their contribution and overall importance for classification. No overlap between the features used by these two classifiers was observed. For the host classifier, we see that more protein features are predictive of a single isolation host type. For the tissue classifier, we see that more protein features are more synergistic in their prediction, being associated to multiple tissue isolation sites.

5.5.1 Host classification

The host classifier using >2000 features performed only slightly better with a *f1_score* of 0.97 versus 0.94 using the most 12 important features (Supplementary material 1B). A slightly lower *f1_score* for the test than training data is to be expected since it is likely that some rarely occurring proteins only occur in the training data and not in the test data. Inspection of the confusion matrix (Supplementary material 1B) revealed one strain isolated from a human and one strain isolated from a *pig-boar* genome to

be wrongly classified. One of these genomes, GCA_001005165, belongs to a zoonotic *Mycoplasma capricolum* isolated from human blood which was misclassified as being isolated from a ruminant. The misclassification as well as the absence of any clear difference in its pathogenesis proteins from other *M. capricolum* strains indicates this zoonotic potential might be the result of opportunism. The other of the two misclassified genome is GCA_000815065 from *M. flocculare*, which was wrongly classified as being isolated from a ruminant.

5.5.2 Tissue classification

The tissue isolation site classifier using all >2000 features performed only slightly better with a *f1_score* of the 0.89 versus an *f1_score* of 0.84 when only using the 12 most important features. Only three genomes (GCA_000319465, GCA_012934855 and GCA_017389835) were misclassified based on the classifier with the 12 most important features. The first genome is from *Mycoplasma haemominutum 'Birmingham 1';* the second genome is from *M. phocoena* infects the urinary tract of harbour porpoise; and the third genome is from an unspecified *Mycoplasma* from the gut microbiome of a buffalo. The first two genomes are from species that only occur 1 time in our dataset while the third genome turned out to be from the gut microbiome and not to be associated to an infection of the gut. Although the three examples above show that tissue isolation site for some genomes of rarely occurring host are misclassified, we do see that the tissue isolation site for other rarely occurring hosts in our dataset, such as one dog (GCA_000238995) and two cats (GCA_000186985, GCA_000200735) infecting *Mycoplasma*, where accurately predicted to be blood infecting. Therefore, we can conclude that tissue classification is to some extent host specific and to some extent non-host specific.

As can be seen in **Figure 20**, for isolation-host classification, *human* and *pig-boar* appear to have a higher similarity in their classification while for isolation-tissue, *blood* and *uri-genital* are closest.The *gastro-intestinal* classification is somewhat more dissimilar, while classification of the *respiratory tract* as isolation site is most dissimilar to the other classes.

5.5.3 Proteins important for predicting host trophism

We further investigated which protein signatures are most important for host and tissue classification for our classifiers based on the 12 most important protein features.

The Pfam [42] domains PF06418 CTP_synth_N as well as PF00117 GATase strongly contribute to predicting human as isolation host class. Surprisingly although very important for correctly predicting human infecting Mycoplasma species, the protein is only present in a few human infecting strains such as M. capricolum (GCA_001005165), Candidatus Mycoplasma girerdii (GCA_002215425), M. penetrans (GCA_000011225, GCA_004127945) and Candidatus M. haemohominis (GCA_008326325). From the 128 genomes that contain these proteins, only the 5 were isolated from humans. Among these are the 2 M. capricolum zoonotic strains and of the four known isolation sites, 4 were isolated from human blood infections. As such, we postulate this protein complex as a requirement for opportunistic infection

of humans through the blood stream. Nucleotide synthesis is known to be critical for growth of bacteria in human blood [448].

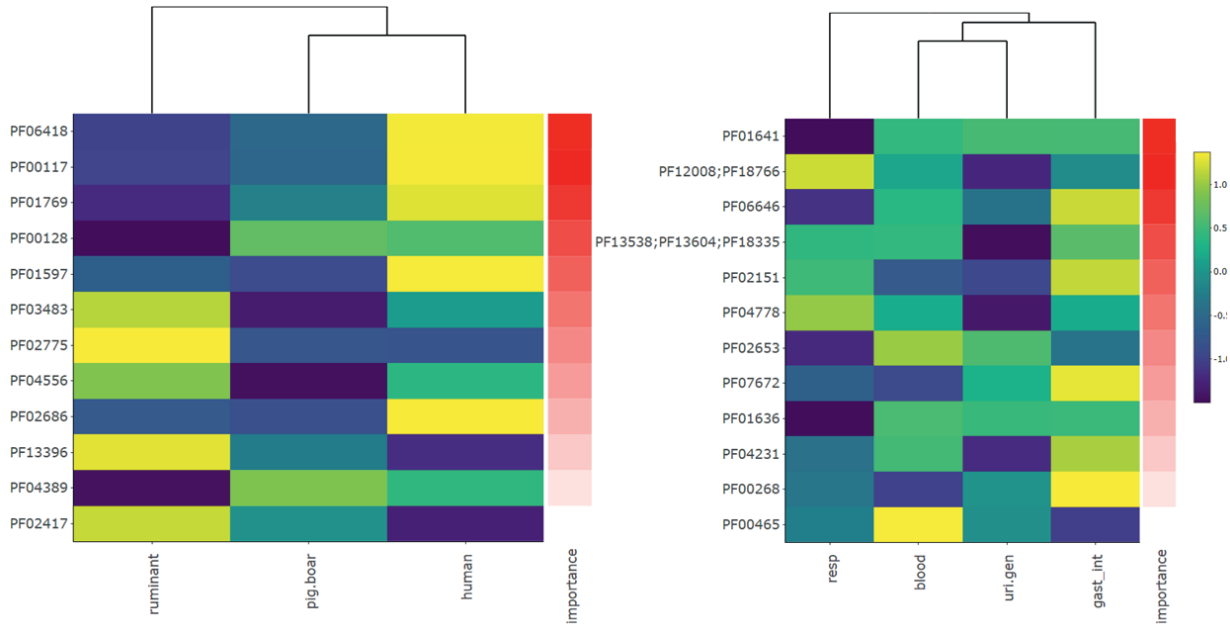

Figure 20. Left host isolation-site classifier feature importance and feature contribution scaled per row. Right tissue isolation-site classifier feature importance and feature contribution scaled per row.

PF01769, the magnesium transporter MgtE, is the third most important feature for host classification. This magnesium transporter has a high contribution to predicting the host class *human* and to lesser extentcontributes to predicting *pig-boar* as host class. Also, this protein is only present in three human infecting Mycoplasma, namely *Candidatus M. haemohominis* (GCA_008326325) and *M. capricolum subsp. Capricolum* (GCA_001005165) while being abundant in *pig-boar* and *ruminant* infecting *Mycoplasmas*. Although KO studies associate this protein to be magnesium transport, it is unknown if this is the primary function of this protein [440]. MgtE is involved in regulating many virulence factors in *Aeromonas hydrophila* as well as in fine tuning regulation of virulence proteins in *Pseudomonas aeruginosa* [449], [450]. PF01597, a glycine cleavage system is another important contributor to predicting human as the isolation host. This protein is however only present in a single human infecting *Mycoplasmas*, namely *M.capricolum subsp. Capricolum* (GCA_001005165) and was reported by Kaminga et al. [368] as a predictor of ruminant and pig infecting *Mycoplasma*. Indeed, PF01597 is much more common in *M. hyopneumoniae* but might contribute to *M. capricolum subsp. Capricolum's* ability to opportunistically infect humans through the blood stream. As can be seen from the above examples, protein features with a high contribution to predicting a single class are not always the most commonly occurring within that class. In the case of *human* infecting *Mycoplasmas,* it appears that the most important protein features are those few proteins that help identify the few zoonotic and opportunistic *Mycoplasma* strains that infect the blood stream.

PF00128 and PF04389 have the highest contribution to predicting pig-boar as isolation host. P00128 is an Alpha-amylase while PF04389 is a Peptidase family M28 protein. PF00128 and PF04389 are both present in present in all M. hyosynoviae strains proteins.

PF03483 and PF02775 have the strongest contribution to predicting *Mycoplasma* that infect ruminants. PF03483 is a B3/B4 domain found in tRNA synthetase beta subunits and other synthetases, while PF02775 is a thiamine pyrophosphate (vitamin B1) binding domain. Also, DpnII restriction endonuclease PF04556 and phospholipase_D-nuclease N-terminal PF13396 strongly contribute to predicting ruminant infecting *Mycoplasma*.

5.5.4 Proteins important for predicting tissue trophism

We further investigated proteins that are most important for tissue classification. The tissue classifiers shows that methionine sulfate reductase A (MetA) PF01641 to be important for urinary tract and genital infections. MetA is a known virulence determinant for M. genitalium which infects the urinary tract while being necessary for proper adhesion [451]. The second most important protein for classification is PF12008; PF18766, a type I restriction endonuclease, which has the highest contribution to predicting the class 'respiratory tract' as isolation site.

PF06646 MG289, an extra cytoplasmic thiamine binding lipoprotein has the highest contribution to predicting the *gastro-intestinal* tract as tissue isolation site, as well as its high contribution predicting *blood* as the tissue isolation site. Additionally, it was shown that MG289 enhances microbial invasion and persistence in *Mycoplasma genitalium* [452].

PF02653, a branched-chain amino acid transport system was found to have the highest contribution to classifying the tissue isolation site *blood* and the tissue isolation sites *uri-genital*. Literature confirms that transport and uptake of branched-chain amino acids is important for protein synthesis and their requirement for environmental adaptation [453]. Other obligatory parasites like the intracellular pathogen *Francisella* lost all branched-chain amino acid biosynthetic pathways and rely on dedicated uptake systems for their survival in the host [454]. Branched-chain amino acids are essential for lymphocyte responsiveness and proper functioning of other immune cells [455], which puts them at the interface of pathogen-host interaction.

5.5.5 Strengths and weaknesses of classification

We repeated the host classification, allowing genomes from rarely occurring hosts isolated from other host to be in the train and testing dataset. Although the accuracy of the training data remained high at 94%, the accuracy for the test data dropped to around 75% and the *f1_score* dropped to 55%. This means that some *Mycoplasma* from host classes contain some of the 12 important features, resulting in misclassification. Classification for strains with rarely occurring hosts is not feasible since there are too few samples for both test and training data.

We also tested if combined prediction of host and tissue was possible to find *host_tissue* specific features. Unfortunately, too few genomes remain when combining tissue and host isolation site information and filtering on a minimum of 5 genomes (Supplementary material 1B). The BioSamples metadata used in our analysis required a manual normalization, combining different metadata fields and normalizing the various labels used in these metadata fields. Therefore, we want to emphasize the importance of 1) sequencing larger numbers of genomes, also for more rare hosts and 2) the importance of well-defined metadata for sequenced genomes to allow for machine learning approaches to predict strain specific properties such as the host isolation and tissue isolation site.

The use of these host and tissue classifiers reveals some of the strengths and weaknesses of classification. For example, those proteins identified as most important for classifications of *human* are those which can best predict the few strains that infect *human* instead of their normal host. Furthermore, because we minimize the number of proteins needed for our prediction, we find proteins that best split species over multiple classes, meaning that many proteins that are identified as important for classification are present in multiple host or tissue types. This makes the approach used in this study less suitable for identifying proteins typical for a single host or tissue type. Furthermore, it should be noted that our classifiers are optimized to predict all available classes equally optimizing for the *f1_score* to balance false positives and false negatives. This is desirable when creating a robust classifier without favouring any single class as intended in our study. However, for medical purposes, such as predicting the risk of a zoonotic outbreak, having maximum recall for *human* infecting *Mycoplasmas* would be desirable since false positives would be preferred over false negatives. Therefore, we advise to train new classifiers when used for such purposes.

5.6 Conclusion

In this study we demonstrated that clustering based on different GO Biological function categories of proteins for *M. pneumoniae*, can provide insight in terms of which functional groups of proteins are most important for pathogenesis. Our study revealed differences in the functional groups of proteins important for *M. pneumoniae* pathogenesis and the pathogenesis of *Mycoplasmas* in general. The GO functional group of proteins *'Interspecies interaction between organisms'* is important for pathogenesis of *Mycoplasmas* in general, while '*Quorum sensing'* and *'Biofilm formation'* proteins are important for *M. pneumoniae* pathogenesis. Furthermore, we show that a small set of proteins can be used to classify host and tissue specificity of various *Mycoplasmas*. Most proteins important for classification were found to have corroborative evidence for their importance to be available in literature, while some might provide new insights such as the proteins identified that differentiate human infecting zoonotic strains from non-zoonotic strains. Finally, our analyses show the feasibility of predicting species properties such as host and tissue types based on genomic information, as well as the importance of high-quality sample meta-data to enable classification through machine learning.

5.7 Funding

This work has been supported by European Union through the FP7 programme under grant agreement No. 305340 (INFECT), the SysteMTb project (HEALTH-F4-2010-241587) and the Horizon 2020 research and innovation programme under grant agreement No. 634942 (MycoSynVac) and from The Netherlands Organisation for Health Research and Development (ZonMW) through the PERMIT project (Personalized Medicine in Infections: from Systems Biomedicine and Immunometabolism to Precision Diagnosis and Stratification Permitting Individualized Therapies, project number 456008002) under the PerMed Joint Transnational call JTC 2018 (Research projects on personalised medicine - smart combination of pre-clinical and clinical research with data and ICT solutions). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

5.8 Author contribution

NZ performed the main analyses and wrote the draft manuscript. MSD and VDMS participated in the design of the study and supervised and directed the research. MSD and VDMS revised the manuscript. All authors contributed to the writing of the final version of the manuscript.

5.9 Supplementary material

Additional file 1AB List of *Mycoplasma* genomes & Random Forest classification of host and tissue isolation data. All Supplementary files, Figures as well as additional code not present in the supplementary files of the published manuscript are available at: https://github.com/NielsZondervan/PhD Thesis.

Chapter 6
Exploring the adaptability and robustness of *M. pneumoniae* central carbon metabolism

Submitted for publication:

Niels A. Zondervan, Eva Yus, Daniel C. Sévin, Sira Martinez, Carolina Gallo, Peter J. Schaap, Maria Lluch-Senar, Luis Serrano, Vitor A. P. Martins dos Santos and Maria Suarez-Diez. "Exploring the adaptability and robustness of *M. pneumoniae* central carbon metabolism". Preprint available online at https://biorxiv.org/cgi/content/short/2022.08.08.503180v1.

6.1 Abstract

In this study we explored the adaptability and robustness of glycolysis and pyruvate metabolism of *Mycoplasma pneumoniae (MPN)*. We used a dual approach; we analysed metabolomics data collected for a large number of OE and KO mutants and perturbation samples. Furthermore, we trained a dynamic model of central carbon metabolism and tested the model's capacity to predict these mutants and perturbation samples as well as identify key controlling factors in central carbon metabolism. Our analysis of metabolite data as well as our model analysis indicate *MPN* metabolism is inherently robust against perturbations due to its network structure. Two key control hubs of central carbon metabolism were identified.

6.2 Introduction

Mycoplasma are gram positive bacteria adapted to an obligatory parasitic lifestyle able to infect a broad range of hosts [456]. It is estimated by the CDC that 2 million infections with *M. pneumoniae (MPN)* occur in the US alone on a yearly basis [457]. These infections lead to conditions ranging from mild to severe respiratory illness including life threatening conditions such as auto-immune diseases [458]. Therefore, it imperative to improve our understanding of *MPN*.

MPN has been established as a model organism for systems biology and large dataset collections are available informing on its genome, transcriptome [459], proteome [460], metabolome [461], and transcripcional adaptations [462]. The metabolism and energetic expenditure of *MPN* have been thoroughly studied by combining a genome-scale constraint based model of metabolism with detailed experimental characterizations [20]. Neither energy production nor uptake of protein building blocks appear to limit growth of *MPN* only protein synthesis was found to be growth limiting [437], [461]. Maintenance requirements are high for *MPN*, so most of the energy is devoted to maintenance instead of growth [20].

MPN like other *Mycoplasma*'s has adopted to its pathogenic lifestyle leading to a severely reduced genome. Despite their small genome size and limited number of enzymes and relatively low number of regulators [463], *Mycoplasma*'s are still able to adapt to a large number of conditions and still genes can be removed, as they have been seen not to be essential [435]. Many essential genes are however only present in some *Mycoplasma* species, suggesting alternatives to a minimal genome exist [368]. Since there are only few regulatory elements in the genome [463] we hypothesized that a lot of the adaptability of *MPN* to adapt to changing growth conditions must be due to network structure and to allosteric control of its metabolism.

In this study we investigated the metabolism of *MPN* with a focus on central carbon metabolism and its allosteric control. Dynamic models were successfully used to investigate regulation and adaptation of central carbon metabolism to changing environmental conditions in other organism [464]–[471]. Therefore, in this study we will combine analysis of metabolomics a large number of samples taken from varying environmental conditions, OE and KO mutants with a dynamic model of glycolysis and pyruvate metabolism, to identify key controlling metabolites and enzymes in central

carbon metabolism. We tested single or combined addition of 1) an ATPase reaction, 2) O_2 inhibition of Lactate Dehydrogenase and 3) NAD regeneration by NoxE using O_2 to this model as potential mechanisms for *MPN's* adaptability to various conditions. We trained and tested the model's ability to predict a wide range of environmental conditions, single and double overexpression mutants as well as mutants with single gene deletions. The model was able to predict these mutants with reasonable accuracy. In a recent study, local sensitivity analysis on a dynamic model of *E. coli* central carbon metabolism identified robustness as one of properties of central carbon metabolism of *E. coli* [467]. This robustness is a system property resulting from the many feed-forward and feed-backward interactions in metabolism, such as allosteric control of glucose uptake as well as lactate and acetate metabolism. Another study in *E. coli* revealed that only three metabolites (FBP, F1P and cAMP) account for about 70% of the expression variability of central carbon metabolism enzymes through control of two transcription factors [472]. Similarly, our model predicts the central carbon metabolism of *MPN* to be inherently robust to changing conditions and identifies two main hubs of metabolic control. Clustering of samples of FBA OE and LDH KO mutants corroborate assumed allosteric control of LDH FBP. Additionally, the analysis of metabolomics data of *MPN* indicated that glycolipid metabolism might be linked to the high energy metabolites needed for growth of *MPN*. Our findings are in agreement with recent findings were some key lipids were identified to be needed for *MPN* growth on serum free medium [21].

6.3 Materials and methods

Bacterial strains and culture conditions

M. pneumoniae strain M129 (passage 33-34) was grown in modified Hayflick medium and transformed by electroporation with the pMT85 transposon as previously described [437]. Briefly, cells were split 1:10, and washed twice with 10 mL and collected in 300 µl Electroporation buffer (8 mM Hepes·HCl, 272 mM sucrose, pH 7.4) three days later. Cells (50 µl) were electroporated with 5 µg plasmid in 1 mm gapped cuvettes at 1.25 kV, 100 Ω, 25 µF (Gene Pulser Xcel Electroporator, Bio-Rad). Cells were recovered in Hayflick for 2 h at 37°C, diluted 1:5 in Hayflick with 200 µg mL^{-1} gentamycin, selected for three days and then maintained with 80 µg mL^{-1} gentamycin. The cell lines used are detailed in **Table 14**.

Transposon insertion mutants obtained by haystack mutagenesis

For the isolation of *M. pneumoniae* mutants, we used a collection of strains carrying insertions of transposon Tn4001 [473]. The presence of the desired mutant was assayed by PCR using one primer that hybridizes to the transposon (directed outwards), and a second primer specific for the gene of interest. Mass spectroscopy is used to verify absence of the corresponding protein.

Over Expression mutant construction

Genes to be overexpressed where cloned in the transposon Tn4001 [474] control of the promoter of the EF-tu gene [474].

Growth curves

To obtain equal amounts of each sample, initial inocula for the growth curves were quantified. Briefly, cells were grown for 3 days in 25-cm² flasks, collected in 1 mL medium and 100 µl was used for quantification using the BCA (bicinchoninic acid) protein assay kit (Pierce, see below). Same amounts of total protein (1 µg) were aliquoted per well in a 96-multiwell plate in duplicates. Two hundred µl of Hayflick medium was added per well and the cells were incubated in a Tecan Infinite plate reader at 37°C. The "growth index" (absorbance 430/560 nm, settle time at 300 msec and number of flashes equal to 25) was obtained every hour for 5 days as published [437]. To quantify growth, we determined two slopes of the growth curve. The first one is based on the time interval from 10 to 30 h ("early slope") and the second one on the whole growth curve ("late"). The early slope was determined by considering the maximum median of the slope between two time points (eq. 1) separated by three time measurements over successive periods of 30 time points. The late slope was determined by considering the maximum median value of the slope between two time points separated by four time measurements (eq. 2) over successive periods of 30 time points.

Early Slope = (value (time [i]) – value (time[i+3])] / (time[i]-time[i+3] (eq. 1)

Late Slope = (value (time [i]) – value (time[i+4])] / (time[i]-time[i+4] (eq. 2)

The early slope is more representative of growth, while the late slope reflects the metabolic activity.

On the other hand, biomass was quantified at 48 h (early stationary phase) by inoculating a twin 96-well plates, in the same conditions as above. After incubation for two days at 37°C under, medium was sucked out, cells were carefully washed twice with 200 µl PBS and lysed with 100 µl lysis buffer (10 mM Tris·HCl, 6 mM $MgCl_2$, 1 mM EDTA, 100 mM NaCl, 0.1% Tx-100, pH 8, and 1× Protease Inhibitor Cocktail, Roche) at 4°C. In the same first 96-well plate, cell lysates were kept on ice and extracted protein was quantified by BCA Protein Assay Kit (Pierce, see below).

The protein concentrations at 48 h and early slope are more representatives of growth, while the late slope and the value of A430/560 at midpoint reflect the metabolic activity. These four parameters of growth and metabolism were analysed for each batch of experiments. Outliers (larger than quartile 3, Q3) by at least 1.5 times the interquartile range (IQR), or smaller than Q1 by at least 1.5 times the IQR) were removed to calculate the mean and the standard deviation of each of the parameters for each batch. Values larger or smaller than the mean by at least 2 times the standard deviation of each parameter were considered to determine fast- and slow-growing/metabolizing clones, respectively.

6.3.1 Strain cultivation and growth conditions of mutant and perturbation samples

A 300 cm^2 flask was inoculated 1:10 with the lab stock and 100 mL of Hayflick and grown for 3-4 days at 37C. Then, medium was removed, and cells were scrapped and resuspended in 12 mL medium. From this inoculum, 75 cm^2 flasks were seeded with 1 mL of inoculum in 20 mL of Hayflick. After 6 hours of incubation (i.e. when cells reached stationary growth phase) the cells were treated as follows, before the standard extraction protocol:

Glucose starvation: remove medium and add new Hayflick medium without glucose. Incubate sample for 5 h at 37C. Long incubation time is required to deplete glucose from Hayflick medium.

Amino acid starvation: take half of the medium, add 200 mg of DL-serine hydroxamate (10 mg/mL), mix and add again to the cells; incubate cells with for 15 min at 37C.

Fe^{2+} depletion: Add directly to the flask the iron chelator 2,2′-Bipyridine at a final concentration of 3 mM, incubate for 30 min at 37° C.

Oxidative stress: Add directly to the flask H_2O_2, to 0.5%, incubate 15 min at 37C.

Glycerol: Add directly to the flask glycerol to 1% -v/v, incubate 30 min at 37C.

6.3.2 Sample preparation for metabolomics

M. pneumoniae cells were grown in 6-well culture dishes as described above until reaching 80-90% confluency. Culture medium was aspirated, and cells were rapidly washed twice at 37° C with 1 mL of buffer (75 mM ammonium carbonate at pH 7.4 and 0.1% glucose). After aspiration of washing buffer, plates were immersed in liquid nitrogen to quench metabolism and stored at -80° C for less than 4 days until further processing. After aspiration of washing buffer, plates were immersed in liquid nitrogen to quench metabolism and stored at -80° C for less than 4 days until further processing.
To extract metabolites, plates were placed on a 75° C heating block and 700 µL of extraction solution (70%-v/v ethanol in water at 75° C) were added to each well. After incubating for 3 min, the supernatant was collected and transferred to ice, and the extraction was repeated once. Pooled extracts were dried under vacuum and stored at -80° C prior to metabolomics analyses.

6.3.3 Nontargeted metabolomics

All samples were measured in triplicate. Metabolomics samples were analysed by flow-injection time-of-flight MS with an Agilent 6550 iFunnel QToF instrument (Agilent, Santa Clara, CA, U.S.A.) operated in negative ionization mode at 4 GHz high-resolution in a range from 50-1,000 m/z using published settings [475]. The mobile phase was 60:40 isopropanol:water (v/v) and 1 mM NH4F at pH 9.0 supplemented

with 10 nM hexakis(1H-, 1H-, 3H-tetrafluoropropoxy)phosphazine and 80 nM taurocholic acid for online mass correction. Spectral processing and ion annotation based on accurate mass within 0.001 Da of metabolites in the *M. pneumoniae* MyMPN database [7], allowing for $[M-H]^-$ and $[M+F]^-$ ions and $[1x^{12}C->1x^{13}C]$ neutral gain and keeping for each metabolite only the ion with lowest *m/z* in case of multiple matching ions, was performed using Matlab R2015b (The Mathworks, Nattick, MA, U.S.A.) as described previously [475]. Metabolomics data were normalized to the summed abundance of a group of amino acids (Ser, Pro, Ala, Val, Thr, Leu/Ile, Met, Phe, Tyr) found to strongly correlate in each sample. In *Mycoplasma* amino acids are not made but imported and they are fairly constant. Therefore, we could use the summed values for the less variable amino acids to normalize. A similar approach is used in free label quantitative proteomics. Subsequently, $\log_2$-transformed fold-changes and *P*-values (two-sided *t* tests, with q-values computed from raw p-values to enable false discovery rate adjustment [476]) were calculated to determine relative metabolite abundances compared to control samples and their statistical significance.

6.3.4 Targeted metabolomics

Samples were injected into a Waters Acquity UPLC with a Waters T3 column (150 mm x 2.1 mm x 1.8 mm; Waters Corporation, Milford, MA) coupled to a Thermo TSQ Quantum Ultra triple quadrupole instrument (Thermo Fisher Scientific, Waltham, MA) with electrospray ionization. Compound separation was achieved by a gradient of two mobile phases (i) 10 mM tributylamine, 15 mM acetic acid, 5% (v/v) methanol and (ii) 2-propanol. In total, 138 metabolites covering carbohydrate and energy metabolism, amino acid metabolism, nucleotide metabolism and other pathways were targeted. Further details are published elsewhere [477].

6.3.5 Proteomics

Cells were grown in a 25-cm2 flask for 3 days as above, washed with PBS and lysed/collected in 4% SDS, and 0.1 M Hepes·HCl pH 7.5. Samples were reduced with dithiothreitol (15 µM, 30 min, 56°C), alkylated in the dark with iodoacetamide (180 nmols, 30 min, 25°C) and digested with 3 µg LysC (Wako) O/N at 37°C and then with 3 µg of trypsin (Promega) for eight hours at 37˚C following FASP procedure (Filter-aided sample preparation 48). After digestion, the peptide mix was acidified with formic acid and desalted with a MicroSpin C18 column (The Nest Group, Inc) prior to LC-MS/MS analysis. The peptide mixes were analysed using a LTQ-Orbitrap Velos Pro mass spectrometer (Thermo Fisher Scientific) coupled to an EasyLC (Thermo Fisher Scientific). Peptides were loaded onto the 2-cm Nano Trap column with an inner diameter of 100 µm packed with C18 particles of 5 µm particle size (Thermo Fisher Scientific) and were separated by reversed-phase chromatography using a 25-cm column with an inner diameter of 75 µm, packed with 1.9 µm C18 particles (Nikkyo Technos). Chromatographic gradients started at 93% buffer A and 7% buffer B with a flow rate of 250 nl min-1 for 5 minutes and gradually increased 65% buffer A and 35% buffer B in 120 min. After each analysis, the column was washed for 15 min with 10% buffer A and 90% buffer B. Buffer A: 0.1% formic acid in water. Buffer B: 0.1% formic acid in acetonitrile.

The mass spectrometer was operated in DDA mode and full MS scans with 1 micro scans at resolution of 60.000 were used over a mass range of m/z 350-2,000 with detection in the Orbitrap. Auto gain control (AGC) was set to 1 E6, dynamic exclusion (60 seconds) and charge state filtering disqualifying singly charged peptides was activated. In each cycle of DDA analysis, following each survey scan the top twenty most intense ions with multiple charged ions above a threshold ion count of 5,000 were selected for fragmentation at normalized collision energy of 35%. Fragment ion spectra produced via collision-induced dissociation (CID) were acquired in the Ion Trap, AGC was set to 5e4, isolation window of 2 m/z, activation time of 0.1 ms and maximum injection time of 100 ms was used. All data were acquired with Xcalibur software v2.2.

Proteome Discoverer software suite (v2.0, Thermo Fisher Scientific) and the Mascot search engine (v2.5, Matrix Science were used for peptide identification. Samples were searched against a *M. pneumoniae* database with a list of common contaminants and all the corresponding decoy entries (87,059 entries). Trypsin was chosen as enzyme and a maximum of three mis-cleavages were allowed. Carbamidomethylation (C) was set as a fixed modification, whereas oxidation (M) and acetylation (N-terminal) were used as variable modifications. Searches were performed using a peptide tolerance of 7 ppm, a product ion tolerance of 0.5 Da. Resulting data files were filtered for FDR < 5 %. Protein Top 3 areas were calculated with unique peptides per protein.

6.3.6 Data and model management

All omics data, modelling files as well as a backup of modelling pipeline and simulation outputs are available via the Seek data and model management platform for maximum reproducibility (http://doi.org/10.15490/FAIRDOMHUB.1.INVESTIGATION.133.3) [41], [478]. SBtab [479], a tabular exchange format was used to add minimal information compliant with the Minimal Information Requirements In the Annotation of Models (MIRIAM)[480] compliant annotation and to add Systems Biology Ontology (SBO) identifiers [481] for metabolites, reactions and parameters in the model.

6.3.7 Data analysis

Relative metabolite measurements were log10-transformed following the recommendation of Jauhiainen et al [482]. Pearson correlations between metabolites were computed using R v3.4.2[430]. To remove batch effects in the metabolites measurement, values were normalized by dividing them by the median measured metabolite value per batch [483]. Fold Change (FC) metabolite measurements for the 40 independent samples were analysed through Principal Component analysis using the prcomp package. Pearson correlation between samples and metabolites were calculated and used to generate heatmaps of sample correlations and metabolite correlations. We used the metabolite correlation matrix for [M+F]-ion data and filtered on correlations with at least p-value cut-off of 0.001. We calculated the Euclidean distance using the complete linkage method and used Hierarchical clustering and cut tree to identify 6 clusters in the metabolite correlation data. Absolute measurements needed for simulations with the model were obtained by

multiplying relative metabolite values from [M-H]$^-$ measurements with quantitative metabolite measurements at 24 h. Similarly, enzyme concentrations of samples with overexpression (OE) of enzymes were obtained by multiplying relative measurements for these mutants with absolute measurements of the wild type at their respective time point. These computations were performed using Python. Additionally, we calculated Pearson correlation between metabolite concentration and estimated growth of 24 time-series samples (P-value <0.05) applying Benjamin Hocheman Multiple testing correction.

6.3.8 Model construction and numerical implementation

A dynamic model was built of glycolysis and pyruvate metabolism. A base model containing all reactions in glycolysis, pyruvate metabolism from the MyMPN database [7].Different additions to this model were tested such as individual and combined additions of 1) an ATPase reaction, 2) LDH inhibition by oxygen and 3) a NoxE reaction for NAD regeneration using oxygen. The tested models include i) the base model ii) the base model and the ATPase reaction iii) base model with NoxE reaction iv) base model with both ATPase and NoxE reaction v) based model with NoxE reaction and LDH inhibition by oxygen and vi) the base model with all three modifications. In case intermediate metabolites were not measurable, reactions were lumped in a single reaction. This was the case for Phosphoglycerate kinase (PGK), Glycerate phosphomutase (GMP) and enolase (ENO). These three reactions were combined in reaction re07 lumping the enzymatic reactions of PGK&GMP&ENO. Similarly, phosphotransacetylase (PTA) and acetate kinase (ACK) were combined in reaction re10 which lumps the enzymatic reactions of PTA&ACK. Allosteric control was assumed to be similar to allosteric control in *Lactococcus lactis* as presented in the model by Costa *et al.* [484] due to the lack of *MPN* specific information on allosteric control. Allosteric control includes three activator and five inhibitor effects. Reactions were modelled using modular rate laws except for transport reactions for which Hill type kinetics were used. Enzyme concentrations were included as reaction parameter to allow model predictions at varying protein concentrations. The base model contains 10 equations and 72 parameters of which 10 represent experimentally determined enzyme concentrations, 5 represent equilibrium constants (K_{eq}) and 1 is a Hill coefficient. The remaining 56 parameters represent Michaelis-Menten constants, activation constants and inhibition constants which are not known for *MPN*. The model was built using COPASI [485].

6.3.9 Initial parameter values

Proteomics measurements for 6, 24 and 48-hour timepoints were used as estimates for enzyme concentrations at respective time points. In case of multi-subunit enzymes, the most abundant single copy subunit was chosen to represent the enzyme concentration. Many lower abundance subunits are only expressed under specific conditions and as such are not representative of the abundance of these glycolysis enzymes. We also tested using the average, but no major differences were found. Equilibrium constants were gathered from www.equilibrator.org assuming an ionic strength of 0.1 and a pH of 7 [486]. Initial values for Monod constants and allosteric control constants were randomly selected between 0.01 and 100-fold of the observed

metabolite concentrations at 24h. An overview of the six models, reactions and equations can be found in Supplementary file 1 A.

6.3.10 Model selection and parameter estimation

The base model has 72 parameters of which 56 are unknown while model that includes ATPase, NoxE and O_2 inhibition of LDH has 80 parameters of which 63 are unknown. Parameter estimation was performed training the models on metabolomics steady state data obtained from growth curve samples for 6h, 24h and 48h time points grown on medium containing 60 mM of glucose. These samples were selected as training data due to the completeness of the data available for these three times points. Only for these three samples, measurements for all 17 metabolites present in the model were available. In addition, protein copy number, glucose uptake, lactate secretion, and acetate secretion measurements were available for these three samples. Steady state concentration for 6h, 24h and 48h grown on a lower glucose concentration of 10 mmol were used as internal validation data. Internal validation data is used by COPASI to stop the parameter estimation algorithm from overfitting parameters to the training data. The large time interval between the samples means that metabolite concentrations in each sample can be assumed to be independent from the concentrations of the other samples. Therefore, each sample was treated as an independent steady state. COPASI's [485] build in Genetic programming algorithm was used to estimate parameters using a maximum of 1000 generations with a population size of 500 models with normalized sum of squares as weights. 100 independent parameter estimations were run per model. Optimal parameters were searched within a range of 10^{-2}-10^{2}-fold of the observed metabolite concentration at 24h for Monod constants and allosteric control constants while maximum reaction velocity values were searched within a range of 10^{-2}-10^{3}. The performance of the six models were compared based on the distribution of the mean square error values for each of the 100 parameter estimations.

6.3.11 Simulations, local and global sensitivity analysis

The model was used to predict steady state concentrations for 40 independent samples comprised of OE and knock out (KO) mutants, perturbations, and time-series measurements in different growth conditions measured in triplicate. In these simulations, the input for the model was concentration data of 11 metabolites: acetyl coenzyme A (AcCoA), acetate (ACE), adenosine diphosphate (ADP), adenosine triphosphate (ATP), coenzyme A (CoA), diacylglycerol phosphate (DGP), fructose 6-phosphate (F6P), fructose-1,6-bisphosphatase (FBP), glucose 6-phosphate (G6P), glyceraldehyde-3-phosphate (GAP), lactate (LAC), nicotinamide adenine dinucleotide (NAD), reduced nicotinamide adenine dinucleotide (NADH), phosphoenolpyruvate (PEP), orthophosphate (Pi_Int), pyruvate (PYR), external glucose (GLC_Ext). Measurements for NAD and NADH are approximate. For each sample, 1000 steady state simulations were performed while sampling from the log normal distribution of metabolite measurements. By comparing sampled measurements and sampled simulation values, measurement error and its propagation are incorporated in model predictions.

Not all metabolites present in the model were measured for all independent samples. Reference values from measurements taken at 6h, 24h and 48h of growth on high glucose concentrations used to train the model were used to set the initial concentration of NAD, NADH and orthophosphate. Reference values were also used for CoA, Acetyl-CoA, and Lactic Acid (LAC) for some of the independent samples (Supplementary Material 2).

To compare the error between simulated and measured metabolite concentrations in a consistent manner, we used the symmetric Mean Absolute Percentage Error (sMAPE). sMAPE is a measure of prediction accuracy used for forecasting methods. This method has the advantage of providing an equal error to positive and negative errors for log normal distributed data such as metabolite measurements and predictions [487].

We performed global sensitivity analysis for all kcat, Monod constants, activation and inhibition constants using a 100,000 Latin Hypercube sampling [488], [489]. Samples were constructed by sampling from the log linear distribution of each parameter's respective search range.

The above described operations were performed using Python 3.6.5 with the Tellurium 2.0.18 and Roadrunner 1.4.24 high performance SBML simulation and analysis libraries [490], [491]. The pyDOE package was used for Latin Hypercube sampling. Conda version 4.3.21 was used for package management.

6.3.12 Modelling oxygen diffusion

Oxygen concentrations were calculated based on the initial oxygen concentration in the culture flasks and the acetate production rate which requires NAD to be regenerated from NADH by the oxygen dependent reaction catalysed by NoxE. The initial oxygen concentration was calculated with the ideal gas law, using the temperature used in cultivation (37 degrees Celsius) and atmospheric pressure. Volumes, surface area height of the medium were calculated based on the medium and inoculant volume and the specifications of the T300 cell culture flask [492].

Diffusion of oxygen from the head space into the medium was calculated using the Wilke and Chang correlation [493] while Fick's law [494] was used to calculate the diffusion of oxygen to the bottom of the flask at 6h, 24h, 48h and 96 hours of growth. The calculated oxygen concentrations were added to the metabolomics measurements for the 95 independent samples.

6.4 Results

Datasets were collected growing *MPN* in a large number of conditions. *M. pneumoniae* was grown in suspension until sedimenting after 6 h of incubation in rich medium in non-aerated, non-stirred conditions mimicking its host environment. At several time points during the growth of wild-type *MPN*, samples were taken for metabolomics, biomass, pH and acetate concentration measurements. In addition, relative metabolite concentrations were measured by untargeted metabolomics for 40 different samples from environmental perturbations, genetic mutations and at 6h, 24h, 48h and 96 hours of growth. The metabolomics data and targeted proteomics data for these

samples is available in Supplementary file 2. Among these 40 datasets, there were data corresponding to OE mutants for all glycolysis and pyruvate metabolism enzymes except for pyruvate dehydrogenase (PDH) of which the complex is large to clone and OE, as well as for the KO of LDH (Mpn674). The fold change in mRNA and or protein concentrations for genes targeted in each mutant were measured. Of the 40 datasets, 17 are mutants that target enzymes for which a reaction is present in the model. Of these mutants 2 are KO mutants and 14 are OE mutants and 1 is a combined KO and OE mutant (**Table 14**). Additionally, there are 6 mutants targeting enzymes in in the pentose phosphate pathway which is connected to the glycolysis via F6P.

Changes in mRNA and protein concentration of enzymes targeted in over expression (OE) and knock out (KO) mutants were also measured. **Table 14** gives an overview of these conditions and mutants. Unless stated otherwise, relative metabolite concentrations were measured at steady state in non-aerated conditions. In cases where different conditions were used, or where additional omics data were measured, this is indicated in **Table 14**. Targeted metabolomics were used to measure protein concentration for all OE mutants targeting central carbon metabolism enzymes.

Table 14. Overview of KO, OE mutants, perturbation and time-series samples.

EXPERIMENT	TYPE	ENZYME IN MODEL
24H_TIMECOURSE *	Absolute Metabolomics, Proteomics, Glucose uptake, acetate lactate secretion rate and lactate secretion rate	NA
48H_TIMECOURSE *	Absolute Metabolomics, Proteomics, Glucose uptake, acetate lactate secretion rate and lactate secretion rate	NA
6H_TIMECOURSE *	Absolute Metabolomics, Proteomics, Glucose uptake, acetate lactate secretion rate and lactate secretion rate	NA
24H_TIMECOURSE_4 **	Glucose uptake, acetate lactate secretion rate and lactate secretion rate	NA
48H_TIMECOURSE_4 **	Glucose uptake, acetate lactate secretion rate and lactate secretion rate	NA
6H_TIMECOURSE_4 **	Glucose uptake, acetate and lactate secretion rate	NA
BLANK_CONTROL_7	Control	NA
WATER_CONTROL_7	Control	NA
WT_5	Control	NA
WT_PERTURBATION_7	Control	NA
KO51_MUTANT_6	mutant, glpD KO	NA

MPN025-OE_6	mutant, tsr OE	FBA
MPN025-OE_7	mutant, tsr OE	FBA
MPN051-KO_5	mutant, glpD KO	NA
MPN051-OE_7	mutant, glpD KO	NA
MPN250-OE_5	mutant, pgi OE	PGI
MPN250-OE_7	mutant, pgi OE	PGI
MPN302-OE_6	mutant, pfkA OE	PFK
MPN302-OE_7	mutant, pfkA OE	PFK
MPN303-OE_5	mutant, pyk OE	PYK
MPN303-OE_6	mutant, pyk OE	PYK
MPN303-OE_7	mutant, py OE	PYK
MPN430-OE_6	mutant, gap OE	GAP
MPN606-OE_6	mutant, eno OE	ENO
MPN627-OE_7	mutant, ptsI OE	NA
MPN674-KO, NOXE OE_5	mutant, ldh KO, noxE OE	LDH, NOXE
MPN674-KO_5	mutant, ldh KO	LDH
MPN674-KO_6	mutant, ldh KO	LDH
MPN674-OE_5	mutant, ldh OE	LDH
MPN674-OE_7	mutant, ldh OE	LDH
TN674_MUTANT_7	mutant, ldh KO	LDH
TN051_GLY_PERTURBATION_7	mutant, perturbation glpD KO	NA
AA_PERTURBATION_6	Perturbation	NA
FE_PERTURBATION_6	Perturbation	NA
GLU_PERTURBATION_6	Perturbation	NA
GLUCOSE_STARV_PERTURBATION_7	Perturbation	NA
GLY_CTRL_PERTURBATION_7	Perturbation	NA
GLY_PERTURBATION_6	Perturbation	NA
OX_PERTURBATION_6	Perturbation	NA
WT_NOGLUC_PERTURBATION_7	Perturbation	NA
M129_TIMECOURSE_24H_3 ***	time course, perturbation	NA
M129_TIMECOURSE_48H_3 ***	time course, perturbation	NA
M129_TIMECOURSE_6H_3 ***	time course, perturbation	NA
M129_TIMECOURSE_96H_3 ***	time course, perturbation	NA

* Used to train the model.
** Used to validate the model.
*** *M. pneumoniae M129* grown in aerated conditions.

We explored the measurements of metabolite concentrations for the various samples shown in **Table 14**. Similar conditions as well as KO of genes in the same pathway

cluster together. An example of this is the clustering of all M129 samples which are the only samples grown in aerated conditions (**Figure 21**).

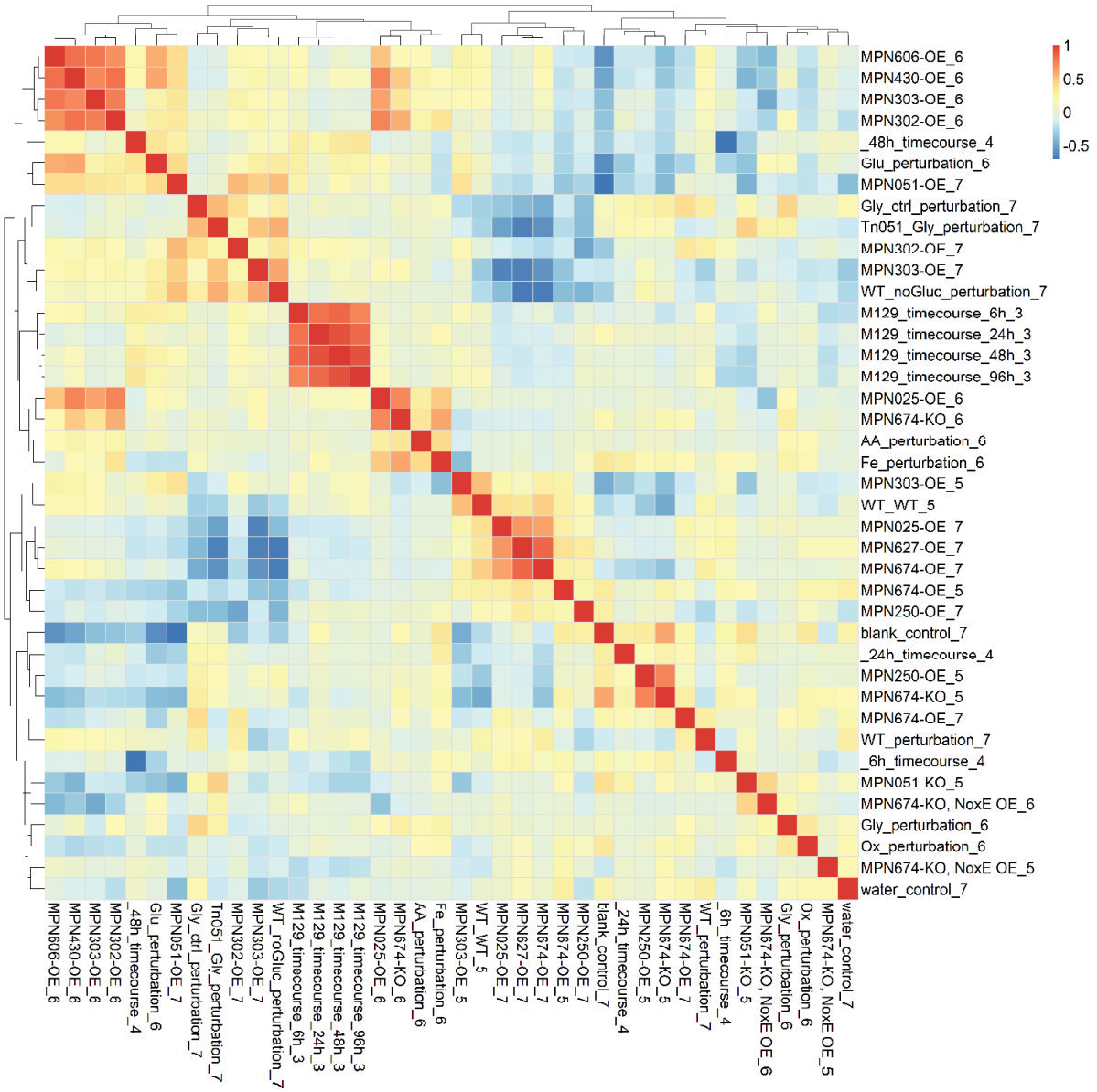

Figure 21. Heatmap of mutant and perturbation sample clustering based on their metabolite profile

Two clusters exist of OE mutants targeting glycolysis. The first cluster contains an OE mutant of FBP, phosphotransferase MPN627 involved in mannitol and mannose uptake [20] and LDH. This clustering corroborates the assumed allosteric activation of LDH by FBP. The second cluster contains OE mutants of ENO, GAP, PYK and PFKA. Another interesting cluster contains *MPN* perturbation, growth without oxygen and growth without amino acids which cluster together with both an LDH KO and FBA OE. These four samples have in common that the conditions are growth inhibiting. The clustering of FBA OE which degrades FBP together with LDH KO can be explained by

the positive allosteric control of FBP on LDH. Two annotation techniques were used to measure metabolite abundance, $[M+F]^-$ ion and $[M-H]^-$ ion detection [495]. Some differences are present in the correlations between individual metabolites in the $[M+F]^-$ ion and $[M-H]^-$ data, however, clustering of samples for both detection techniques is highly similar. A heatmap that compares both $[M+F]^-$ ion and $[M-H]^-$ data can be found in the Supplementary files 1B.

In addition to clustering of samples based on their relative metabolite concentrations we studied the clustering of metabolites of these samples. We use the Pearson correlations between metabolites to build a network of metabolite-metabolite interactions (**Figure 22**).

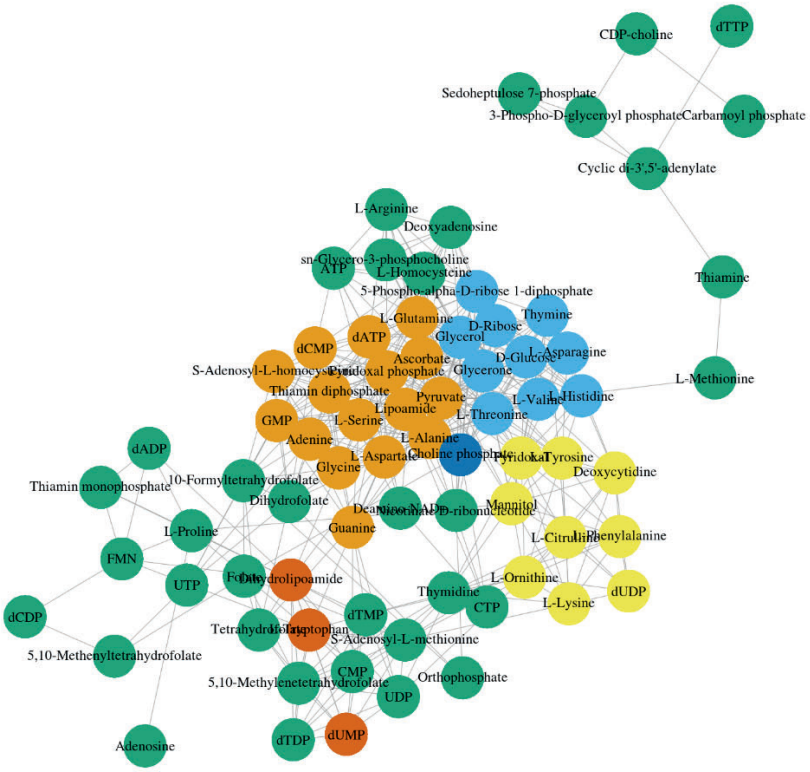

Figure 22. Metabolite correlation network based the Pearson correlation of metabolite measurements of the 40 independent samples. The 6 clusters were assigned based on hierarchical clustering of the correlation data.

We identified a neatly defined cluster structure. The largest cluster contains sn-glycero-3-phosphocholine, CDP-Choline, folate and methionine cycle metabolites, orthophosphate, all nucleotide three phosphates (ATP, CTP, UTP) and pentose phosphate metabolism (PPP) metabolites. This cluster suggest a link between sn-glycero-3-phosphocholine catabolism, energy production, and nucleobase salvaging by phosphorylation and de-oxidation. Higher concentrations of PRRP associated with

these metabolites are needed to convert nucleobases into ribonucleotides, while increase in ATP is needed for phosphorylation of deoxyribonucleotides:

$$Nucleobase + PRPP <=> Nucleotid - monophosphate + PPi$$

In humans, phosphatidylcholine (lecithin) is used as main nutritional source for one carbon metabolism (CH_3) [496] One carbon metabolism and its relation to nucleotide synthesis in cancer cells has been extensively studied [495]. It has been argued that in human cancer cells, glycolysis can produce enough energy for growth by diverting its flux to other metabolic pathways including one-carbon metabolism. Indeed, several reactions of one-carbon metabolism contribute to ATP and NADPH production. Similarly, it has been argued that phosphatidyl choline plays a major role in the nutrition of *MPN* as it is by far the most abundant available carbon source in the lungs and for these reasons is also used as carbon and nitrogen source by pathogens like *P. aeruginosa* [497], [498] and claimed to be used as carbon source by *MPN* [499].

The second largest cluster contains alanine, aspartate, glutamate, serine, pyridoxal phosphate (vitamin B6), S-adenosyl-L-homocysteine, glycine, lipoamide, pyruvate as well as adenine, guanine, GMP, dATP and dCTP. Pyruvate is positively correlated with dATP and dCTP which are needed for DNA synthesis. This cluster is strongly associated to the sub cluster of sn-Glycero-3-phosphocholine and tetrahydrofolate metabolites from the first cluster.

Based on the observed cluster, there are three main lessons to be learned. Firstly, clustering of sn-Glycero-3-phosphocholine with ATP, CTP and UTP metabolite correlation profiles supports the theory [497] that glycerol-3-phosphate derived from sn-Glycero-3-phosphocholine functions as a carbon and energy source for *MPN*. Addition of phosphatidylcholine to a defined minimal medium for *MPN* indeed optimizes growth [21]. Secondly, sn-Glycero-3-phosphocholine clusters together with folate and methionine cycle one carbon metabolites and as such is likely the main one carbon donor in *MPN* metabolism. Thirdly, positive correlation between CDP-choline, 3-phospho-D-glyceroyl-phosphate and sedoheptulose-7 phosphate, as well as between sn-Glycero-3-phosphocholine and the cluster containing PRPP indicate a link between sn-Glycero-3-phosphocholine and pentose phosphate metabolism.

6.4.1 Over Expression of glycolytic enzymes

To further study the control of different glycolytic enzymes on central carbon metabolism we, analysed the fold change of enzymes in glycolysis and pyruvate metabolism when OE single as well as some combinations of glycolytic enzymes **(Table 15**).

Table 15. Log2 Fold change expression values of OE mutants. I: Mutant PTA ACK did not show any OE of ACK. II: OE are significantly different from the wild type. III: OE values are significant, but the changes are noisy.

		PTA ACK[I]	*PTA*	*PFK*	*PFK PTA*	*PFK LDH*	*LDH*
		MPN428; MPN53	MPN428	MPN302;	MPN302; MPN428	MPN302; MPN674	MPN674
		fold change	fold change	fold change	fold change	fold change	fold change
PTS	MPN207	0.07	0.18	-0.04	-0.18	-0.07	0.10
PFK	MPN302	-0.08	-0.10	2.73 [II]	2.66 [II]	2.71 [II]	-0.20
PGI	MPN250	-0.12	0.09	-0.16	-0.05	-0.04	0.04
FBA	MPN025	0.28	0.54	0.04	0.47	0.75	0.62
GAPDH	MPN430	-0.22	-0.14	-0.27	-0.31	0.19	0.07
PGK	MPN429	-0.01	0.04	0.07	0.20	0.05	-0.01
PGM	MPN628	0.06	0.21	-0.20	0.03	0.51	0.25
ENO	MPN606	0.24	0.30	0.13	0.32	0.41	0.35
PYK	MPN303	-0.20	-0.13	-0.11	0.12	0.01	-0.15
LDH	MPN674	0.37	0.54	0.01	0.34	1.50 [II]	1.69 [II]
lplA	MPN389	0.29	0.18	0.17	0.27	0.35	0.19
pdhD	MPN390	0.03	-0.04	0.08	0.19	-0.08	-0.11
pdhC	MPN391	0.24	0.12	0.10	0.23	0.15	0.15
pdhB	MPN392	0.34	0.19	0.27	0.29	0.37	0.36
pdhA	MPN393	0.07	-0.01	-0.11	0.06	-0.08	-0.09
nox	MPN394	0.71	0.50	0.12	0.43	0.80	0.52
pta	MPN428	2.11 [II]	2.22 [II]	0.28	1.93 [II]	0.26	0.33
ack	MPN533	0.36	0.34	0.28	0.43	0.40	0.30
GlpD	MPN051	0.41	0.44	0.19	0.27	0.60	0.49
atpC	MPN597	-0.16	0.21	-0.21	-0.11	0.27	0.31
atpD	MPN598	0.15	0.29	0.02	0.34	0.53	0.42
atpG	MPN599	0.20	0.30	0.20	0.11	0.71	0.32
atpA	MPN600	0.00	0.16	-0.01	0.09	0.16	0.27

atpH	MPN601	-0.15	0.18	-0.09	0.05	0.43	0.12
atpF	MPN602	-0.08	0.06	-0.30	-0.49	0.18	0.23
atpE	MPN603	-1.53 [III]	-0.52	-0.16	1.23 [III]	0.30	-1.76
atpB	MPN604	0.20	-0.15	-0.54	-0.60	-0.10	0.19
tkl	MPN082	0.02	-0.01	0.03	0.18	0.20	0.26
tim	MPN629	0.21	0.31	0.13	0.13	0.37	0.47

What we see is that when OE these other enzymes in the pathway don not widely change. These results suggests that allosteric regulation and circuit topology might play a great role on the control of central carbon metabolism of *MPN*.

6.4.2 Exploration of metabolite's concentration at steady state: study of associations

We build a dynamic model of central carbon metabolism including glycolysis and pyruvate metabolism. We trained this model with a limited subset of data and use the model to identify key regulatory elements in glycolysis. Different additions to this model were tested such as individual and combined additions of 1) an ATPase reaction to account for varying ATP demand, 2) LDH inhibition by oxygen and 3) a NoxE reaction for NAD regeneration using oxygen. The tested models include i) the base model ii) the base model and the ATPase reaction iii) base model with NoxE reaction iv) base model with both ATPase and NoxE reaction v) based model with NoxE reaction and LDH inhibition by oxygen and vi) the base model with all three modifications.

We found the model with addition of NoxE to have the best fitting in multiple parameter estimations, therefore we kept the addition of the NoxE to the model we used for further simulations/

In case intermediate metabolites were not measurable, reactions were lumped in a single reaction. This was the case for Phosphoglycerate kinase (PGK), Glycerate phosphomutase (GMP) and enolase (ENO). These three reactions were combined in reaction re07 lumping the enzymatic reactions of PGK&GMP&ENO. Similarly, phosphotransacetylase (PTA) and acetate kinase (ACK) were combined in reaction re10 which lumps the enzymatic reactions of PTA&ACK. Allosteric control was assumed to be similar to allosteric control in *Lactococcus lactis* as presented in the model by Costa *et al* [484] due to the lack of *MPN* specific information on allosteric control. Allosteric control includes three activator and five inhibitor effects. Reactions were modelled using modular rate laws except for transport reactions for which Hill type kinetics were used. Enzyme concentrations were included as reaction parameter to allow model predictions at varying protein concentrations. The base model contains 10 equations and 72 parameters of which 10represent experimentally determined enzyme concentrations, 5 represent equilibrium constants (K_{eq}) and 1 is a Hill coefficient. The remaining 56 parameters

represent Michaelis-Menten constants, activation constants and inhibition constants which are not known for *MPN*. The model was built using COPASI [485].

An overview of the model's reactions can be found in **Figure 23**, and the reactions, equations, and the additions tested can be found in Supplementary material 1A.

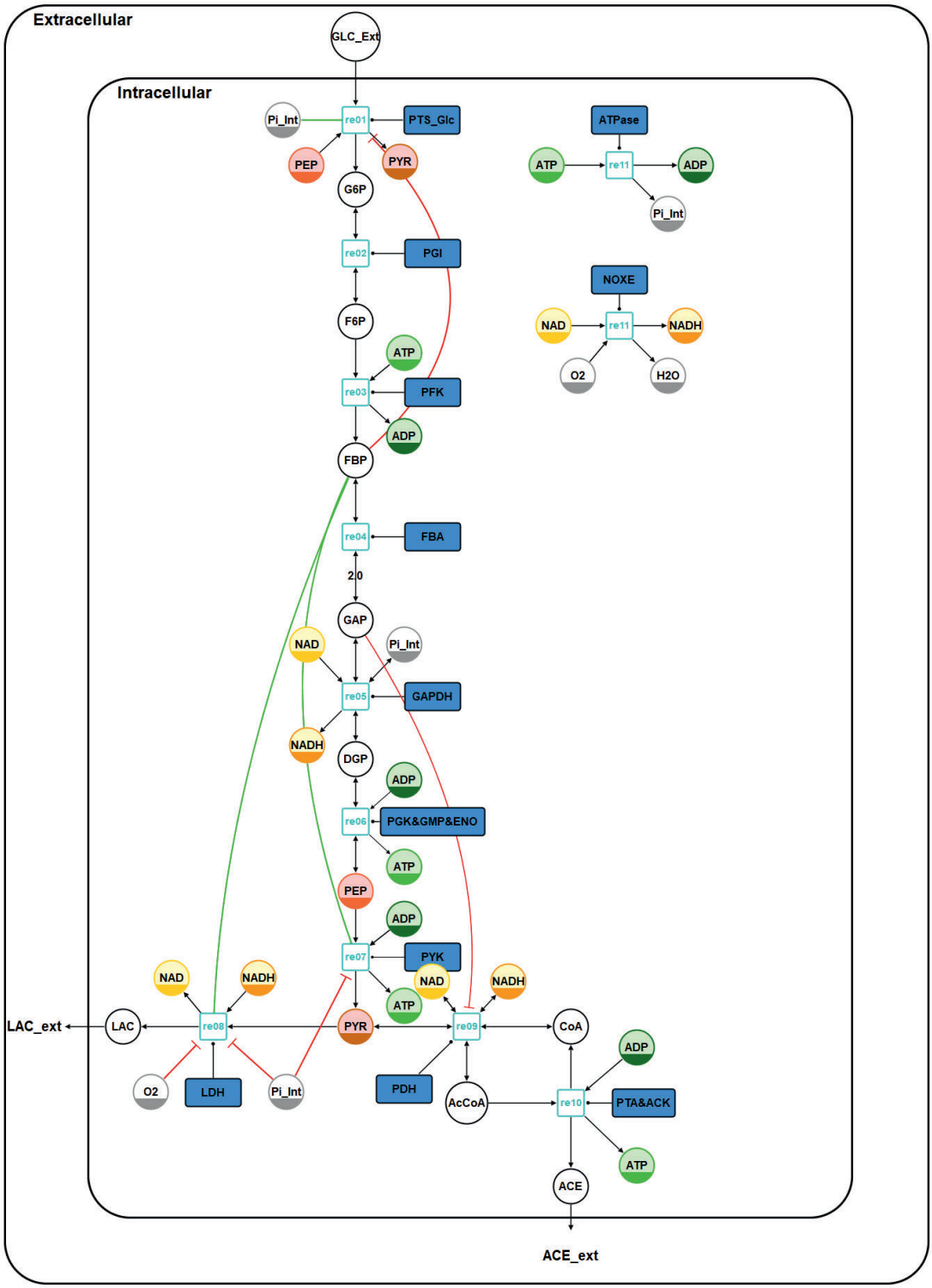

Figure 23. Schema of the model. The diagram meets the Systems Biology Graphical Notation (SBGN) standard [500] *with the exception of the LAC and ACE export arrows which in the model are present as syncs for LAC and ACE: Arrowheads represent reactions (black arrowhead end), catalysis (open circle end), activation (green) and inhibition (red). Circles indicate metabolites. Half-filled circles are clone makers to indicate the metabolites appears multiple times in the diagram (green=adenine nucleotides, yellow=redox equivalents, red=PEP/PYR). Blue filled rectangles describe macromolecules (enzymes, transporters) that catalyse a reaction. In case the reaction*

stoichiometry is different to 1, the reaction stoichiometry is given as text at the reaction arrow. Blue empty rectangles indicate reaction identifiers in the model. Metabolite abbreviations: AcCoA = acetyl coenzyme A, ACE = acetate, ADP = adenosine diphosphate, ATP = adenosine triphosphate, CoA = coenzyme A, DGP = diacylglycerol phosphate, F6P = fructose 6-phosphate, FBP = fructose-1,6-bisphosphatase, G6P = glucose 6-phosphate, GAP = glyceraldehyde-3-phosphate, LAC = lactate, NAD = nicotinamide adenine dinucleotide, NADH = reduced nicotinamide adenine dinucleotide, PEP = phosphoenolpyruvate, Pi_Int = orthophosphate, PYR = pyruvate (PYR), GLC_Ext = external glucose.

We compared the mean square error, for all parameter sets (**Figure 24**). Models that include ATPase (models 2, 4 and 6) have on average the largest error and took the most iterations to reach a stable solution. The addition of NoxE was shown to reduce the error in model predictions (see **Figure 24**). A zoomed in version of **Figure 24** as well as correlation analysis of parameter sets is available in Supplementary file 1B. Since the model with addition of NoxE performed best, further analyses were continued using this model. This model loaded with the best performing parameter set was deposited in BioModels [501] and assigned the identifier MODEL1911200003.

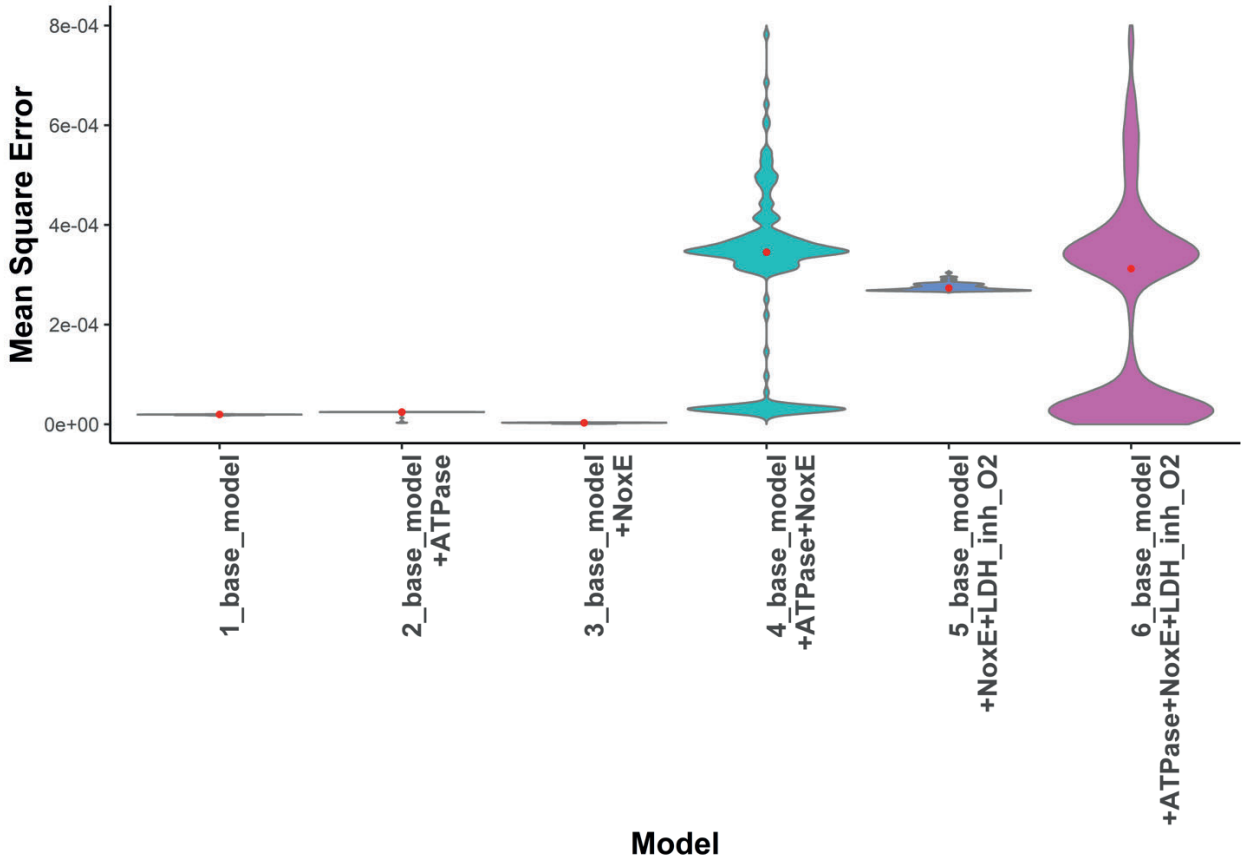

Figure 24. Violin plot comparing the mean square error, for all 100 parameter sets for six models. The red dot indicates the median value.

6.4.3 Simulating perturbations, KO and OE mutants

We used the trained model to predict steady state metabolite concentrations for 40 independent samples. Sample's metabolite concentration mean and standard deviation values were determined by measuring samples in triplicate. In addition to being measured in triplicates, biological replicates were available for all OE targeting glycolysis and pyruvate metabolism.

For each of the 40 samples, 1000 independent simulations were performed with slightly different initial concentrations to explore the impact of biological variability and uncertainty associated to measurement error rates. On average the model predicted these samples with reasonable accuracy (**Figure 25**).

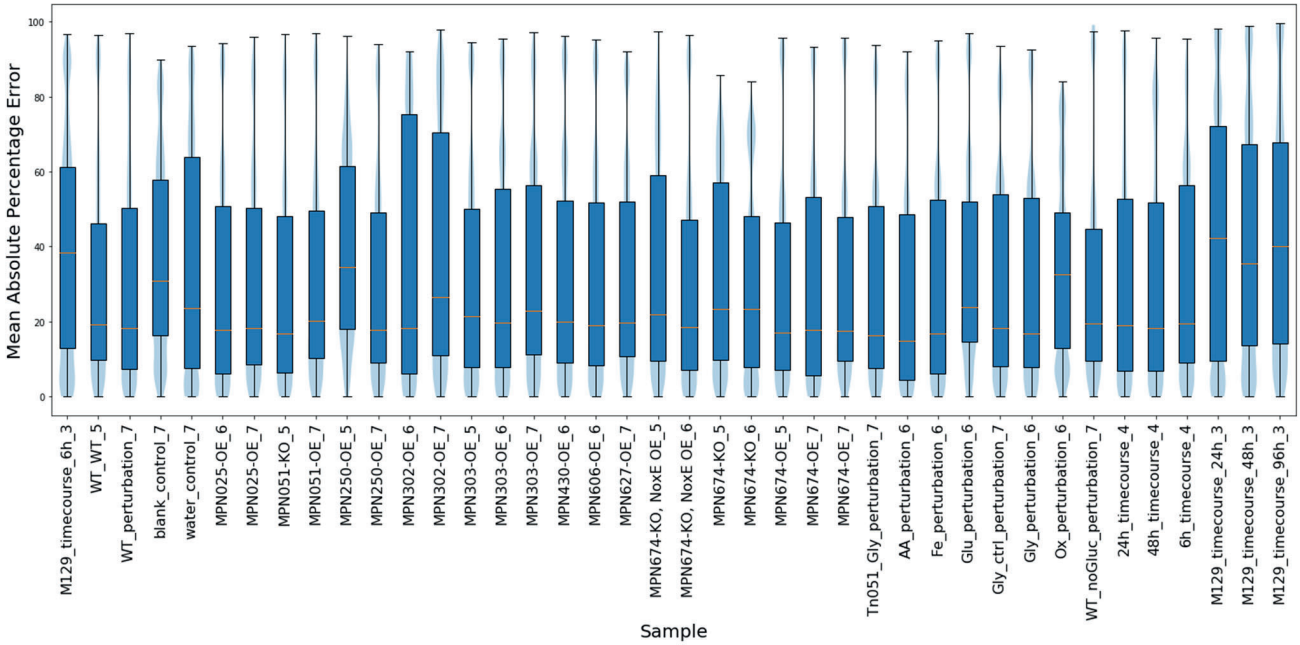

Figure 25. Symmetric Mean Absolute Percentage Error between simulated and measured values using a 1000x times sampling. sMAPE for all metabolites per sample are combined.

The largest error in predicted metabolite concentrations occur for the samples of MPN129 growth in aerated conditions. These results were to be expected as these samples are clearly forming an outlier since the growth conditions are so vastly different from the other samples. From samples corresponding to perturbations in growth condition, the oxidative stress perturbation (0.15% H_2O_2) had the largest prediction error.

The relatively simple model here presented reproduces the states attained under a broad range of perturbations such as glucose concentrations up to a factor 10 lower as compared to the training data. The use of proteomics data is most likely one of the main reasons the model simulates OE and KO mutants of enzymes in glycolysis relatively well. The calculated metabolite concentrations are, on average within a factor 3 of measured values for all mutants and perturbation samples, which is

comparable to the accuracy of the training set. An example of measured and simulated values of sampled simulations can be seen in **Figure 26**.

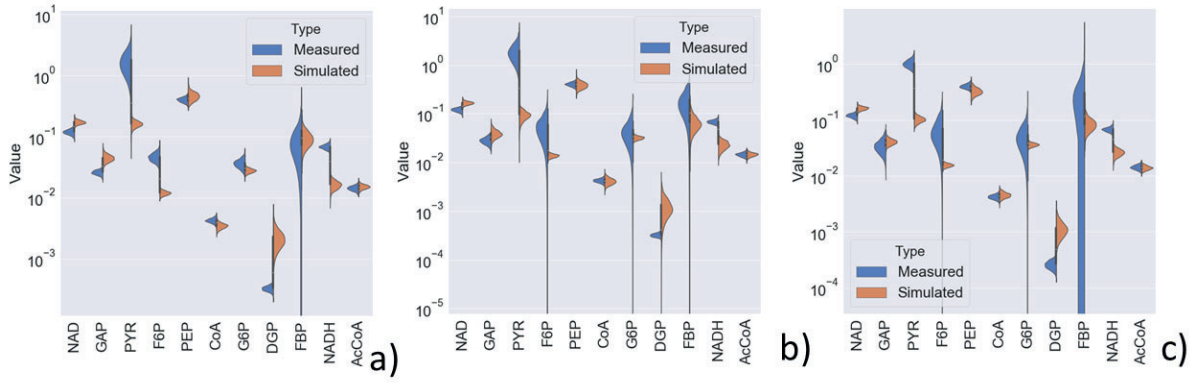

Figure 26. Measured and simulated steady state concentrations at a) 6h of growth, b) 24h of growth and c) 48h of growth on low glucose concentration. Measurement for DGP, NAD and NADH are reference values from samples grown on high glucose concentration.

Additionally, concentrations of for most metabolites are predicted within the 95% confidence interval of the *in-vitro* data (Supplementary material 1C). Exceptions to this rule are PYR and ATP and NADH. Simulated concentrations for ATP and NADH are systematically lower than measured values while PYR is systematically predicted to have a higher than measured concentration. Pyruvate is hard to quantify since it easily is degraded causing variation in its measurement due to for example different times between measurement and sampling. As such we cannot quantify the accuracy of the predictions for pyruvate. For the metabolites such as DGP, NAD and NADH, reference values from high glucose conditions were used to set initial concentrations. As such deviation from these reference values in the simulations in steady state was to be expected.

6.4.4 Metabolic control

We used two types of metabolic control analysis to understand which enzymes and metabolites effort the greatest control on glycolysis: Local Sensitivity analysis and Global Sensitivity analysis. Control coefficients are unit less measures of the relative steady state change in a system variable, in our case the flux through PFK, in response to a relative change in a parameters value. Global sensitivity provides information on parameters that exert control independent of a specific parameter value. We achieved this by sampling parameter sets uniformly from the parameter search space. These parameters sets are not specific for *MPN* since they are not fitted to any *MPN* data. Local sensitivity analysis on the other hand is based on control coefficients derived at the steady state using parameter sets fitted to *MPN* specific data. As such the approaches are fundamentally different and complementary in the information they provide. For our local sensitivity analysis, we use the best performing10 parameter sets to calculate metabolic control coefficients. Since these parameter sets are independent of one another, if metabolic control is higher for certain parameters based on multiple parameter sets, we can conclude the control to be relevant since it is a result of the

fitting to *MPN* data. Additionally, we also performed local sensitivity analysis while sampling from the measurement distribution of metabolites to investigate the effect of measurement error on control coefficients. The steady state changes for each sampled simulation, as such we can see how the uncertainty in metabolite concentrations propagates and creates uncertainty in the control coefficients calculated at these steady states. An overview of our metabolic control analyses can be found in Supplementary file 1D. We found two main control hubs PTS_Glc + PFK and LDH + PDH + PYK. The first control hub consists of parameters associated to PTS_Glc and PFK and represents metabolism in the upper part of glycolysis, the second hub consists of parameter associated to LDH, PDH and PYK and part of pyruvate metabolism.

6.4.5 Simulating combined OE and KO mutants

We used the model to simulate the combined effect of genetic perturbations targeting glycolysis enzymes (OE, KO) combined with a second perturbation, either genetic or environmental. This analysis can identify bottlenecks in central carbon metabolism consisting of combinations of enzymes. Such bottlenecks cannot be identified through local sensitivity analysis or when simulating single over expressions. The expected variations in the flux through glycolysis is shown in **Figure 27**. For most of these combined perturbations, only minor changes were observed. However, simulations of OE of PFK show greatly increased flux through glycolysis while oxygen stress, iron limitation and growth of MPN129 in aerated conditions lead to greatly reduced flux. However, combination of PFK OE with OE of lactate dehydrogenase (LDH) or phosphotransacetylase (PTA) with acetate kinase (ACK), increase the flux nearly as much and is realistic to obtain *in vivo* since OE mutants for each of these enzymes individually are available. Based on these simulations, combined OE mutants were suggested for lab validation: PFK+LDH, PFK+NOXE. Growth curves, protein and metabolite concentrations were measured for the PFK+LDH OE mutant but not for combined PFK+NOXE OE. Combined PFK+PDH OE did not increase flux through glycolysis. However, some positive epistasis on the growth for the combined OE of PFK+LDH was observed, with higher biomass and higher acidification than the individual mutants. None of the combined OE mutants increased the growth rate of *MPN* with respect to the wild type as could eb expected since energy metabolism is not growth limiting in *MPN*[437].

The simulation results of the double OE mutants show that glycolysis in robust. Meaning that the network structure that includes feed forward and feed backward allosteric control, makes the glycolysis of *MPN* inherently robust.

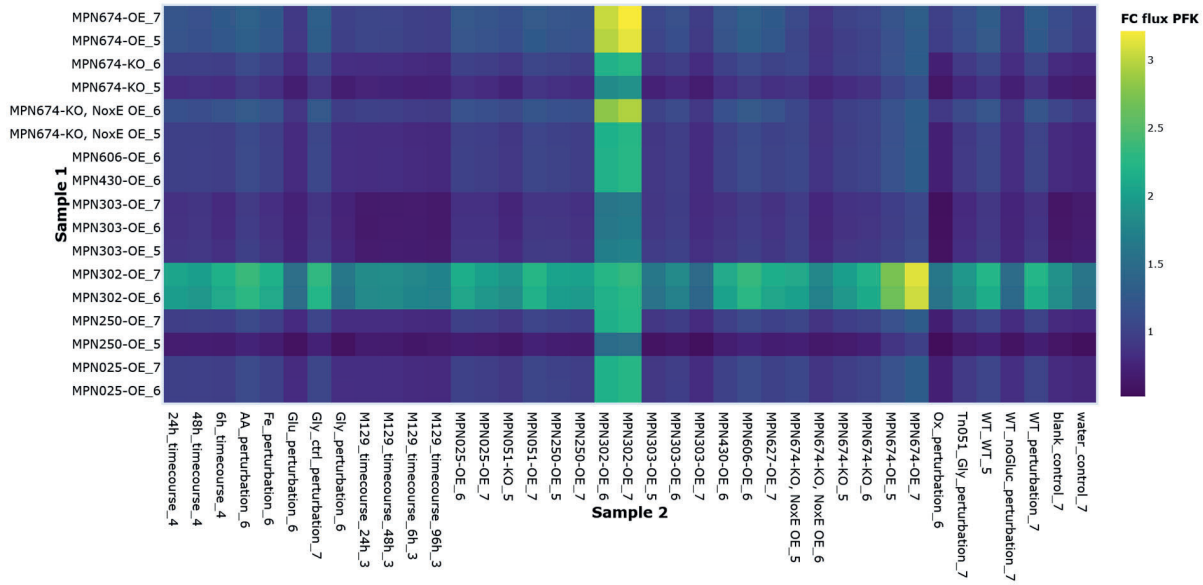

Figure 27. Fold change flux through PFK for combinatorial mutants and perturbations in steady state, relative to the wild type at 24h. The primary sample, of which only the protein OE values are used, are shown on the y-axis while the secondary sample, of which both changes to parameters and metabolite concentrations are used, are shown on the x-axis. The colour represents the fold change in flux through glycolysis compared to the wild type.

6.4.6 Discussion

Metabolomics data for a large number of perturbations, OE and KO mutants were collected. We observe that metabolite concentration of samples in general do not vary that much and that samples of similar conditions as well samples of enzymes OE and KO mutants of enzymes close to other in the metabolic network cluster together. In glycolysis we observed no clear clustering between metabolites. We also see that there is relatively little difference in the expression of glycolytic enzymes when OE enzymes in glycolysis and pyruvate metabolism. These results suggest that central carbon metabolism of *MPN* is robust against perturbation. Literature research revealed that glycolysis and pyruvate metabolism in many species is observed to be robust against perturbations [467], [502], [503]. Arguably robustness is an even more important property for a minimal organism such as *MPN* where biological noise can be expected to have much larger effect than for many organisms with a larger cell volume.

The property of robustness agrees with model results that show only minor changes occur even when OE multiples glycolysis enzymes in silico. We tested the effect of adding a reaction for additions of 1) ATPase, 2) O2 inhibition by LDH and 3) a NoxE reaction. Only the the addition of addition of a NoxE reaction resulted in a much better fit to the training data. Arguably, addition of NoxE improves parameter identifiability since the models without this reaction use a fixed reference values for NAD and NADH. The models that include NoxE allow NAD and NADH concentrations to change, this added flexibility and improves parameter identifiability since NAD and NADH associated parameter can now be used to account for differences in metabolite

concentration opposed to being fixed values based on reference values. Additionally, NAD/NADH are known to have control over central carbon metabolism in other organisms such as *L. lactis* [504], [505] therefore we argue it is likely they also have control over central carbon metabolism in *MPN*.

Steady state simulations showed the model to be flexible since it can predict metabolite concentrations well for all mutant samples. Part of this flexibility is the result of including parameters representing the enzyme concentration, therefore genetic perturbations are accounted for in simulations. Similarly, by using measurements of cofactors from these conditions, the model can approximate the effects of these simulations on central carbon metabolism. Additionally, we argue the flexibility of the model might partly be the result of the inherent robustness of central carbon metabolism in *MPN*.

6.5 Conclusion

In this study we integrated experimental data and model simulations and analysis to explore the robustness of central carbon metabolism of *MPN* in steady state. Firstly, we analyzed metabolomics data. We observed that samples from similar conditions as well as samples of OE and KO mutants of enzymes close to each other in the metabolic network cluster together. Samples from vastly varying conditions, such as aerated conditions, do not cluster with the other samples. Secondly, we build a model to simulate samples of various single or combined perturbations. The simple model presented in this study can predict metabolite concentration with reasonable accuracy for a wide range of conditions and OE and KO mutants. Two control hubs were identified using our dynamic model a) upper glycolysis (PTS_Glc + PFK) and b) lower pyruvate metabolism (LDH+PDH+PYK). No single or combined OE mutant of glycolysis and pyruvate metabolism enzymes resulted in a higher growth rate although OE of PFK and LDH resulted in somewhat higher acidification indicating there might be higher flux through glycolysis. These results are in in agreement with studies that show that glycolysis and pyruvate metabolism in *MPN* is not growth limiting. Both the results from the analysis of our samples as well as the model results, suggest robustness to be a central property of *MPN* glycolysis and pyruvate metabolism.

6.6 Acknowledgement

We acknowledge support of the Spanish Ministry of Economy, Industry and Competitiveness (MEIC) to the EMBL partnership, the Spanish Ministry of Economy and Competitiveness, 'Centro de Excelencia Severo Ochoa 2013-2017', the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program under agreement No 670216 (MYCOCHASSIS) , the European Union's Horizon 2020 research and innovation programme under grant agreement No 634942 (MycoSynVac) , the CERCA Programme / Generalitat de Catalunya, FEDER project from Instituto Carlos III (ISCIII, Acción Estratégica en Salud 2016) (reference CP16/00094) and "Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya" (2014SGR678). We would like to thank dr.ir. A. Rinzema for his expert advice on calculating the various oxygen diffusion rates in our experimental setup.

6.7 Contributions

LS, VMdS and MSD conceived the study. NZ developed the model and performed the simulations with guidance and input from PS, VMdS, LS and MSD. EY, DS, SM and CG performed the experiments with guidance and input from ML-S and LS. NZ wrote the manuscript with input from VMdS, PS, ML-S, LS and MSD.

6.8 Competing interests

The authors declare no Competing Financial or Non-Financial Interests.

6.9 Corresponding authors

Correspondence to Luis Serrano or Maria Suarez Diez.

6.10 Supplementary material

All Supplementary files, Figures as well as additional code not present in the supplementary files of the published manuscript are available at: https://github.com/NielsZondervan/PhD Thesis.

Chapter 7
General Discussion

7.1 Introduction

The objective of this chapter is to reflect on the research performed within this thesis. Firstly, I briefly discuss key aspects of biology of pathogens, addressing similarities and differences between pathogens as well as possible directions for future research. Secondly, I discuss what worked and what did not work well in terms of methodology.

Critical feedback and lessons learned from research are in general underrepresented in scientific literature [506], [507]. However, failures and bottlenecks encountered in research are invaluable to shape the direction of future research and to avoid future researchers repeating the same mistakes their predecessors made. I discuss current developments in methodology and what I think it would be required to bring Systems Biology and modelling to the next level. I will try to answer the overarching research question and sub questions of this thesis:

"What are the patterns in pathogenesis host interaction? "

1. *What are the strategies used by different pathogens to cause illness?*
2. *What are the strategies a model organism like* Mycobacterium tuberculosis *deploys to infect the host?*
3. *How do functional groups of proteins associate to differences in pathogen host interaction?*
4. *Which genes confer zoonotic ability to bacteria?*
5. *which genes determine the host and tissue specificity of bacterial pathogens?*
6. *What are the properties of* Mycoplasma pneumoniae *central carbon metabolism to adapt to different environmental conditions?*

7.2 What are the patterns in bacterial pathogen host interaction?

Before I try to answer the above questions, let us discuss the taxonomy of the various pathogens discussed in this thesis. I discussed three virulence strategies of *Mycobacterium tuberculosis* in **Chapter 2,** identification of regulatory binding motifs associated to these three virulence strategies in **Chapter 3***,* pathogenesis of *Staphylococcus & Streptococcus* species in **Chapter 4***,* predicted tissue and host specificity of *Mycoplasma* species in **Chapter 5** and presented a dynamic model of *Mycoplasma pneumonia* metabolism in **Chapter 6**. The taxonomic relation of these species is visualised in **Figure 28**. Although all species and species groups discussed contain pathogenic bacteria, these pathogens are vastly different from a taxonomic point of view. The closest taxonomic link is that they all belong to the clad of *Terrabacteria,* a taxon that contains two thirds of all prokaryotic life forms.

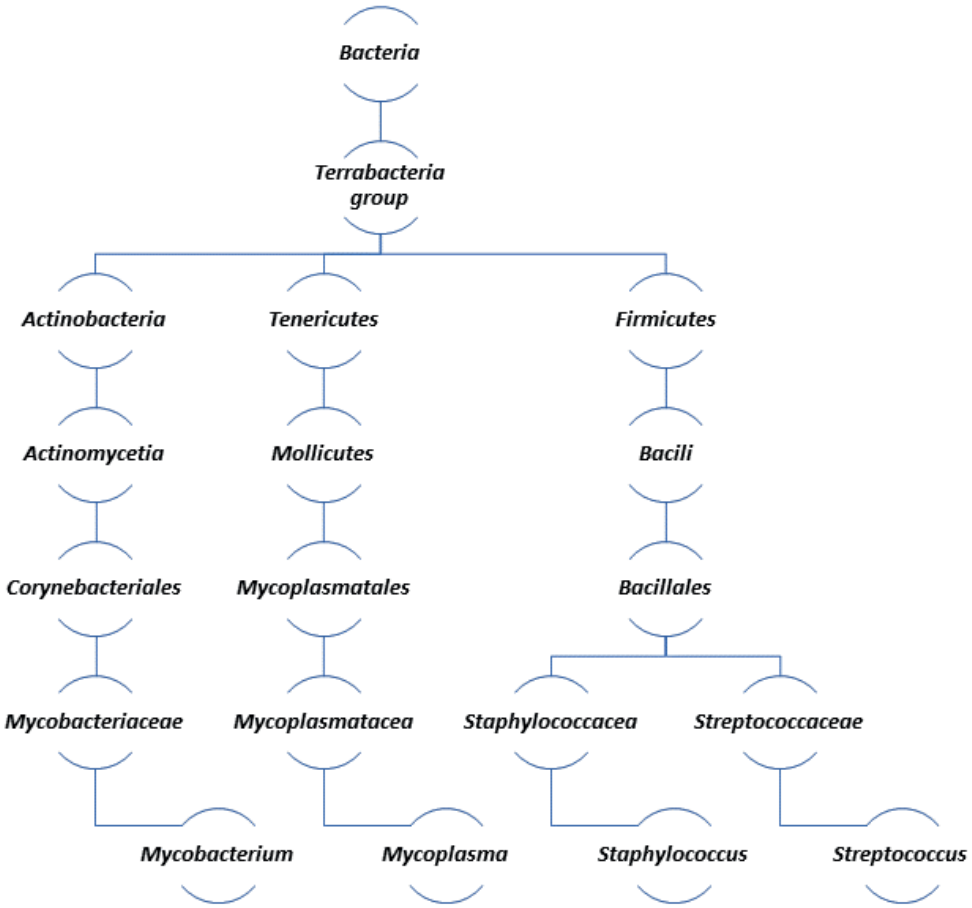

Figure 28. Taxonomic tree of the pathogens studied within this thesis.

The *Terrabacteria* are, as the name suggests, adapted to live on the surface of the Earth. This clad contains bacteria that are resilient against many environmental stresses. This includes resilience against hostile conditions encountered when infecting an animal or human host.

From the organisms discussed in this thesis only bacteria belonging to the *Staphylococcus* and *Streptococcus* genus are closely related, and therefore discussed together in **Chapter 4**. All other pathogens are very distant relatives. However, despite these large taxonomic distances, these pathogens share commonalities in their virulence strategies, the environmental cues that regulated these virulence strategies, and the pathogenic proteins involved in virulence. In the coming paragraphs I will highlight some of these similarities as well as some differences.

7.2.1 Modulation of the host, cell surface and lipid uptake

Staphylococci and *Streptococci* are Gram positive bacteria. Although they evolved from Gram positive bacteria, both *Mycobacterium* and *Mycoplasma* bacteria do not have a regular cell wall or cell membrane. *M. tuberculosis* has a complex cell wall with an outer capsule containing myco-lipids, glycans, glycol-lipids, trehalose monomycolate and dimycolate (TMM, TDM), phthiocerol dimycocerosate (PDIM), pentacyl trehalose (PAT), sulpholipid-1 (SL-1) and diacyl trehalose (DAT) as well as some ESX-1 secreted proteins [86], [508]. This unusual waxy capsule serves many purposes such as immune modulation [86] as well as protection against environmental stresses [509]. Although genetically vastly different, *Mycoplasmas* also adapted to a pathogenic lifestyle by adjusting their surface to their human host. *M. pneumonia* mimics the host cell membrane by incorporating lipids such as cholesterol, sphingomyelin, phosphatidylglycerol and phosphatidylcholines [21], [510][511], [512] from the host and by transferring various sugar groups from the host proteins [513] and lipids [514] to its own membrane. Without the lipids from its human host, *M. pneumonia* cannot survive for long since it lacks the metabolic capacity to produce these lipids by itself [436], [441]. Like *M. pneumonia, M. tuberculosis* takes up cholesterol from its human host [280]. Cholesterol is necessary for *M. tuberculosis* survival during long term infection since its degrades cholesterol for building blocks and energy consumption [515]. In summary, both *M. tuberculosis* and *M. pneumonia* infect the lungs of humans, mimic the hosts cell surface to modulate and evade the immune response and opt for a slowly growing to keep a low profile. Both bacteria have adjusted to the nutritional availability in the host and use host lipids for degradation or as building blocks. Also *S. aureus* uses lipids from the host for its own membrane synthesis during infection, although it does not use these lipids to mimic the host cell surface [516]. Cholesterol being one of the most abundant lipids in human lung cells, appears to be a preferred target of many pathogens [517], especially lung pathogens [518] such as *M. tuberculosis* [97].

7.2.2 Dormancy

Apart from these commonalities, *M. tuberculosis* and *M. pneumonia* are completely different in their infection strategy. *M. pneumonia* has a minimal genome and survives mainly through stealth and evasion while *M. tuberculosis* actively steers towards uptake by macrophages, escape the phagosome or try to survive by steering the immune response towards dormancy and granuloma formation, especially foamy cholesterol rich granulomas [84], [519]. In its dormant state, *M. tuberculosis* can survive for decades while being immune to most drugs, making it such a difficult pathogen to eradicate [9], [11]. Also various *Staphylococci* species have been implicated in granuloma formation and dormancy [516]. Similar to *M. tuberculosis,* the oxidative stress response from the host immune cells upon phagocytosis can induce *S. aureus* species to enter a dormant state [520] from which it can resuscitate when a milder oxidative stress response is measured [521]. There is even a possible *S. aureus* resuscitating infection after 65 years of dormancy [522]. For *Streptococcus* species, the marine pathogen *S. parauberis* has been reported to be able to enter a dormant state when starving [523]. It should be noted though that this dormant state

was reported upon starvation conditions outside of the host, not within the host as is the case for *M. tuberculosis* and *S. aureus*. Both *S. aureus* and *M. tuberculosis* can survive in anaerobic conditions which is essential for survival in phagosomes, although *M. tuberculosis* must enter a near metabolic inactive dormant state to survive in anaerobic conditions.

7.2.3 Survival and escape from the phagosome

All pathogens studied in this thesis modulate the immune response of the host in one way or another. While *M. tuberculosis* tries to buy time by delaying and downplaying the macrophage response, it also actively steers towards being take up by macrophages. *S. aureus* also steers towards uptake by phagocytes, actively using the immune cells to hide and replicate until it breaks out and lyses the immune cell [524], [525]. Various pore forming proteins assist in survival in the phagosome and escape to the cytosol of various cell types [525]. Especially cystic fibrosis trachea and long epithelial cells which are rich in cholesterol [526] are vulnerable to *S. aureus* entry and escape [527].

S. aureus, *S. pneumonia, M. tuberculosis* and in lesser extend *M. pneumonia* utilize toxins to kill human macrophages. *M. tuberculosis* produces Necrotizing Toxin (TNT) secreted by ESX-4, that can induce necrosis in macrophages by depleting NAD [528]. TNT works in concert with pore forming proteins ESX-A and EXB secreted by ESX-1 for permeability of the phagosomal membrane of which the complex regulatory cascade has been discussed in **Chapters 2** and **3**. Although at the time of publishing this information was still relatively controversial, recent studies indeed confirm parts of this regulatory cascade such as the secretion of EsxA-EsxB, dissociation, and subsequent pore formation to be important for modulating the macrophages immune response and subsequent escape of *M. tuberculosis* from the phagosome to the cytosol [529]. Recent years have produced unsurmountable evidence that *M. marinum*, as well as *M. tuberculosis* can escape from the phagosome to the cytosol [530]–[532].

Like *M. tuberculosis, S. aureus* can survive and actively induce phagocytosis by macrophages. *S. aureus* produces pore forming toxin called α-hemolysin toxin Hla as well as leucocidins pore forming toxins [533]. Like *M. tuberculosis, S. aureus* contains four Type VII secretion system and secretes *EsxA-EsxB* toxins that were shown to mediate pore formation in lipids and contribute to meningitis development [533]. *S. pneumonia* also uses a repertoire of toxins such as haemolysins, proteases, superantigens and other agents [364], [533], [534]. Unlike *M. tuberculosis* and *S. aureus, Streptococci* cannot survive within phagosomes. *Streptococci* are mostly facultative anaerobes while some are obligate anaerobes. Some *Streptococci* like *S. pyogenes,* produce toxins that can lead to toxic shock syndrome [412], [535]. In addition to being used to kill immune cells. Like *S. aureus, S. pneumonia* secretes a pore forming toxin, pneumolysin [534]. Both *S. pneumonia* pneumolysin and EsxA specifically target cholesterol rich membranes such as the membranes of lung cells, and phagosomes [234]. Pore formation in *Mtb* [536], as well as activation of AB-toxins and other pore forming toxins in general utilise the acidic conditions in the phagosome as trigger for dissociation and pore formation [537]. Even on a molecular level there are similarities between *S. aureus* and *M. tuberculosis*. Both bacteria use ESX-1

secreted EsxA-EsxB pore forming proteins to gain cytosolic access, which is need to modulate the immune response that leads up to escape from the phagosome [196], [288]. For both bacteria EsxA pore formation was shown to happen at the low pH of around 4 encountered in a maturing phagosome [234], [536]. Type VII secretion systems like ESX-1 and secretion of EsxA-EsxB like proteins appear to be a common pattern in pathogens. Not only is this highly efficient secretion and pore forming complex important for *Mycobacteria* and *Staphylococci* as discussed above, but it also present in *Listeria monocytogenes* [538], *Streptococcus suis* [196] and *Streptococcus gallolyticus* [196], *Salmonella,* and *Yersinia pestis*. Also similarities exist in their regulation, such as regulation by PhoP/Q/R like systems in *Yersinia pestis* [449], *M. tuberculosis* [131], [226] and *Salmonella* [539].

7.2.4 Patterns in bacterial pathogenicity

Multiple bacterial pathogens use the same strategies and same molecular building blocks, as we can see from the discussion on three virulence strategies a) modulation of the host b) survival and escape from the phagosome and c) dormancy.

These examples are an answer to our main research questions:

"*What are the patterns in bacterial pathogenicity?"*

More patterns are likely to be found when exploring different aspects of pathogenesis. For example, in **Chapter 6** we identified metabolic enzymes of *M. pneumoniae* to be the most important predictors of host and tissue specificity. Although we did not further explore this angel further, it is likely that similar metabolic adaptations to these host and tissue types exist in other pathogens.

In **Chapters 2** and **3** we answered the question:

"1. *What are the strategies a model organism like M. tuberculosis deploys to infect the host?"*

We identified immune modulation, phagosome escape and dormancy as the three main strategies to infect the human host. As we can see from our discussion, many of these strategies are deployed by various pathogens. Therefore, we can state that indeed *M. tuberculosis* is a great model pathogen since knowledge of these strategies and their molecular building blocks can in many cases provide insight in the working of other human pathogens. The regulatory cascade of *M. tuberculosis* to switch between virulence strategies is intricately linked with the environmental cues and divalent metal availability in the host environment. The host reduces availability of divalent metals as well as uses toxic accumulation to kill bacterial pathogens [540]. Many bacteria use the restricted availability of iron as a signal to upregulate virulence proteins [115].

In **Chapters 4** and **5** we answer the question:

"2. *How do functional groups of proteins associate to differences in pathogen host interaction?"*

We show that by using guilt by association to predict GO Biological functional Groups of protein annotation, we can find functional groups of proteins that reveal alternative

clustering of strains that coincides with pathogenic properties such as a) zoonotic potential and in lesser extend b) associated to host and tissue trophism.

In **Chapter 2** we discuss the role of divalent metals in pathogenesis, answering the question:

"What are the strategies a model organism like M. tuberculosis deploys to infect the host?"

In **Chapters 4** and **5** we identified protein domain fingerprints associated to zoonotic ability as well as to specific host and tissue types. To answer the research questions:

"*Which genes confer zoonotic ability to bacteria?"*

For *S. suis* we identified two cytolysin proteins to be the main two factors' predictors of zoonotic ability. For *S. agalactiae* we found genes coding for proteins involved in immune evasions such as extracellular nucleases and a protease that breaks down chemokines, Fibrinogen binding protein and three pilus forming proteins to be the main predictors of zoonotic ability.

In **Chapter 5** we also answer the question:

"Can we predict host and tissue specificity of bacterial pathogens?"

Our results show that we can with a can predict three *Mycoplasma* host classes with an F-score of 0.91 while we can predict tissue specificity with an F-score of 0.88. In summary, it is possible with reasonable precision and recall predicting host and tissue specificity for *Mycoplasma's*. In **Chapter 5** we also tried and failed to predict combined host and tissue specificity. This is not surprising since the number of samples we have available for these combinations of tissue and host classes is insufficient.

In **Chapter 6** we try to answer the questions:

"What are the properties of M. pneumonia central carbon metabolism to adapt to different environmental conditions?"

We identified robustness as the main emerging systems property of *M. pneumoniae*. Latin-Hypercube sampling of the parameter space shows that robustness is in general as systems property of central carbon metabolism. To find robustness in a minimal organism such as *M. pneumoniae* which has a small volume with few copies of proteins, and a limited cell surface is not surprising. Two key metabolic hubs in metabolism were identified. The first control hub consists of parameters associated to PTS Glc and PFK representing the upper part of glycolysis, the second hub consists of parameter associated to LDH, PDH and PYK which are part of pyruvate metabolism.

7.2.5 Systems medicine approaches to target bacterial pathogens using phages and CRISPR CAS

One interesting angle in research would be to compare similar strategies such as pore formation in various bacteria that infect humans, not only the molecules that that create these pores, but the regulatory cascades that precede them as well. There are vast similarities in their strategies, such as immune modulation, preference for cholesterol rich membranes, steering towards phagosome uptake, granuloma formation, cytosolic access using pores, and escape to the cytosol. Similarities in environmental cues, regulatory cascades and virulence proteins are present across multiple species. Pooling the knowledge available on various pathogens, should increase our understanding of the intricate interactions, similarities and differences between pathogens and their virulence systems and strategies. This might especially become important with the increased prevalence of drug resistant bacteria. Systems approach to find weakness in pathogenesis strategies and regulatory cascades can lead to systems medicine approaches to combat pathogens. Although such approaches are used, they are still limited and mostly only rely on literature search. Extending such approaches to multiple omics data, regulatory elements, environmental cues and using automation, would be a logical next step.

Additionally, I argue that more 'adaptive medicines' such as the use of phage treatments are essential to keep up the arms race with pathogens. I hypothesize that the combination of phage treatments with the use of CRISPR CAS systems in pathogens, might facilitate many systems medicine approaches. Phage treatment and system medicine approaches can be used as an alternative or in complement with antibiotic treatments [518], [541], [542]. Although CRISPR CAS is normally used by bacteria to protect against bacteriophages, CRISPR CAS can potentially be hijacked to fight bacteria by using phages as vector to deliver CRIPR CAS systems with antimicrobials with bactericidal activity [543]. Phages have been reported as successful vectors for CRISPR CAS deliveries [543]. Examples of such systems medicine approaches could be a) to trigger death of pathogenic bacteria by letting it silence some of its own genes, b) activating dormant *M. tuberculosis* to make them susceptible to drugs and clearance by the immune response or c) silencing essential virulence proteins using the innate CRISPR CAS systems to stop necrotising fasciitis. The given examples are hypothetical, they might not work or have serious complications as well as ethical implications. For example, an estimate 13 million US citizens carry latent Tuberculosis [544], and an estimate 33% of the African population have latent Tuberculosis [545]. However, using phage therapies might become the only option to treat extreme drug resistant bacteria. In 2021 there were 9 clinical trials underway to treat bacterial infections as well as some successful trials with phage cocktail therapies to treat infections with *P. aeruginosa* or *Acinetobacter baumannii* [543]. Experimental treatment of multi drug resistant *M. tuberculosis* with single phage therapy in 2022 showed a favourable clinical outcome in 11 out of 20 patients with no negative side effects. Although further development of these techniques is needed, these first results are promising for patients with extreme drug resistant *M. tuberculosis* that cannot be treated effectively using available antibiotics. The adaptive ability to overcome the bacterial immune system makes phages essential

to supplement the dwindling pool of effective antibiotics. It should be noted that many obstacles still exist in phage treatment, such as difficulty with regulation and quick adaptation and clearance of phages by the hosts immune response [542]. Despite these limitations, the promises of phage therapies due to a) their adaptive nature b) their specificity, and c) their potential to be used as vector for drug delivery or CRISPR CAS systems with bactericidal activity, make them an attractive solution to treat multi and extreme drug resistant bacteria.

7.3 Methodological strengths and limitations

7.3.1 Networks and clustering

Network-based approaches and clustering of omics data is the most extensively used method within this thesis. It also is, arguable, the most effective and time efficient methodology. In **Chapter 2** I used DIVA[52], a precursor of SyNDI [51], a synchronous network data integration framework combined with literature research to create a visual mapping of proteins, environmental signals and regulators that orchestrate the three virulence strategies in *M. tuberculosis*. This methodology was proven to be effective although the use of these tools and methodology might work best with organisms with an abundance of omics data such as is the case for *M. tuberculosis*. The visual model provided in the supplementary file of **Chapter 2** would ideally be a part of a multiple network view opposed to being a non-interactive image. Multiple omics data are presented to a scientist as unordered networks, either as aggregated network or as individual networks by tools such as SyNDI [530]. These unordered networks can be used to create ordered networks/maps to help a scientist understand the biology. Examples of ordered networks are maps of metabolic reactions that are visualised in accordance with reference maps of metabolism such as Escher [546] and MOST-visualization [547] that display model data on well-known map layouts such as the Roche Applied Science 'Biochemical Pathway's wall chart [53], or modular visualisation as used by Pathway Tools [54]. Model databases such and BiGG Model database incorporate Escher. By using the same layout, differences between models can be spotted much more easily. Another example of an order network is the modular map of the three virulence strategies we presented in **Chapter 2** were we ordered proteins and metabolites in modules which in turn were ordered based on the environmental cues encountered in the phagosome**.** The initial intent of our visual model was to provide an ordered network, a map of map of virulence containing multiple layers of omics data. I made attempts to translate this visual map into an interactive map using WikiPathway [548], [549] and Pathvisio [294]. This author believes there to be great merits to such an approach, using direct linkage of omics data, annotation to multiple network visualities including both unordered networks for exploration and ordered modular networks maps that best represent the biology. Ordered modular network maps are maps where nodes such as genes, proteins or metabolites are ordered in pathways and modules and arranged in accordance with reference layouts such as we commonly see in maps of metabolism. Examples would be genome scale metabolic (GEM) models and their maps in Escher [546] or Gene regulatory network models in WikiPathway [549]. Although in theory this approach is nice, conversion of the *M. tuberculosis* virulence map to an interactive

map was at the time of writing not feasible due to several reasons. Firstly, the available tools were too limited, time consuming, buggy, and poor in scaling to make such approaches work for a large manually made model. This is not too surprising since the graphical virulence map already required 16 GB or RAM for editing without containing any metadata or links to other omics data. In general, a lot can be improved in linking data to biological maps. Secondly, building such a map requires well defined standards in identifiers and data storage which at the time of writing were still rather undeveloped. I will discuss bottlenecks, current developments and future developments that might solve these bottlenecks in the section

Breaking the barrier between models and data at the end of this discussion.

Apart from the multiple network visualisations used in **Chapters 2 and 3**, network visualisation and clustering were used for discovery and classification of bacterial traits. In **Chapters 4 and 5** I used clustering-based protein presence and absence belonging to specific GO Biological functional groups [550] retrieved from the GODM database [551], to identify sub-populations of bacteria associated to specific phenotypes in heatmaps, dendrograms, PCA and t-SNE plots. Clustering based on a few selected biological functional groups of proteins was shown to be useful in identifying clusters of strains with different phenotypic traits. This methodology should be transferable to the study of any bacteria within a short taxonomic distance where there are different phenotypic properties observed. The GO functional groups were manually selected based on their suspected importance for the traits to be classified. Alternatively, it could be possible to automatically scan for separation of properties in all GO biological functional groups of proteins, although it should be noted that separation might occasionally occur at random due to the large number of GO biological functional groups of proteins. Having a) more diverse genomes and more phenotype metadata available b) restricting the depth of the GO functional groups searched and c) manual validation of the involvement of a GO functional group in associated traits, can be used to overcome this problem of randomly occurring segregation of strains with different phenotypic traits.

In **Chapter 6** I used network graphs and heatmaps based on metabolomics data to find correlations between metabolites over conditions. This analysis led to some interesting leads into which metabolites might be linked to growth and high energy state.

In **Chapters 4** and **5** clustering led to clearly identifiable sub-groups of strains with different phenotypes. Clustering based on GO functional groups for *Mycoplasmas* was somewhat less clear. Probably because we try to predict multiple tissue and host types, as well as because of overlap in protein features between classes. Classification of *Mycoplasmas* was synergistic in nature and contained many proteins part of core functionalities such as metabolism. From a biological point of view this is not surprising since *M. pneumonia* pathogenesis is an emergent property that incorporates proteins involved in adherence, immune evasion, inflammation, cytotoxicity, gliding motility as well as metabolic adaptation to its host [440]. Hence, any single biological functional groups of proteins do not fully capture the adaptation to any host or tissue type. Pathology of *M. pneumonia* includes adhesion damage, toxic damage, invasive damage, disruption of membrane fusion, nutritional depletion,

inflammatory damage and damaged caused by the immune system [16]. Furthermore, in **Chapter 5** I predict host and tissue specificity over multiple species opposed to within a specific specie as was the case in **Chapter 4**. The results of our study show that *Staphylococci* and *Streptococci* ability to infect humans originate from the presence of a few toxins and pathogenic proteins unlike *Mycoplasma's* and *M. tuberculosis* which uses combinations of proteins belonging to various GO functional groups to adapt to its host. For future studies, it would be interesting to apply the same methodology of using GO biological functional groups of interest to identify clusters of strains with different phenotypes for *Mycobacterial* species. If such future research would be possible largely depends on the availability of strain phenotype data, such as the isolation host and tissue type.

7.3.2 Pattern recognition and clustering

Pattern recognition can be loosely defined as the clustering and of patterns. Classification is the assigning of labels to a to a pattern or cluster. In the case of supervised learning, the label given to classes are known and machine learning models is trained to predict these classes for un-classified data or to classify a 'test data' set which was not used in the training of the classifier. In unsupervised learning, classes are automatically detected and assigned, and classes represent an abstract entity. An example of unsupervised learning would be automatic detection of clusters. An example of semi supervised learning is the identification of binding motifs in upstream binding sites which we deployed in **Chapter 3**. We start with a seed cluster with the known label as belonging to DevR, but we allow the cluster of genes to expand and the pattern predictive of this class to emerge. The approach we used in chapter 3 could be made unsupervised. I propose the following approach to automated motif searches and segregation. Currently, there are two manual steps in our approach: a) selection of genes matching a motif or cluster based on the cut-off p-value for the Fimo motif search and b) negative selection of genes with a known regulatory binding motif for a next iteration. The cut-off p-value determines how stringent the iterative approach will be in adding new genes found by Fimo in the next motif identification round by Meme. This iterative approach can be continued until he input for motif building by Meme matches the genes above the cut-off value by Fimo or if a fixed number of iterations has been performed. The final set of genes can be stored as a group together with the found motif. Optimal p-values for segregation can be chosen in many ways. There are many algorithms available to automate cluster selection such as K-means, Spectral, clustering, DBSCAN, Partitional Clustering Algorithm based on Nearest Neighbours Heuristics [552], Local Density with Glowworm Swarm Optimization [553], projection to latent structures discriminant analysis [554], extended Minimum Spanning Tree [555]. Silhouette score or Rand index combined with hyperparameter searches can potentially be used to benchmark the performance of different clustering algorithms with various setting. By no means would the choice and implementation and fine tuning of automatic clustering algorithms be trivial. However, it would be worthwhile to automate since it would allow scaling up of the methodology to identify all motifs and their putative regulons in a network. Showing the overlap in motifs and regulation can be essential to unravel the complex regulation of genes such as virulence genes in *M. tuberculosis* as we show in **Chapter 1** and in this chapter. The second manual step

in our approach is the negative selection of genes with a known regulatory motif. There are many ways to automate this negative selection. One solution would be to use a second a less restrictive seed phase where more clusters, and as such small sub-clusters are detected. These small clusters can be used to identify less frequently occurring motifs that are present in larger clusters. Another approach would be to select all genes in a cluster that are below the cut-off p-value of a Fimo motif search and use these as seed to find new motifs within larger clusters. The last step would be to store all motif hits and motif patterns and display these different clusters with different colours in SyNDI. Motifs can be compared to motifs from motif databases and can be used to select groups of likely co-regulated genes within SyNDI. Upstream regions of genes can be visualised within SyNDI to show the different putative regulatory binding site regions and their estimated strength of binding represented by the p-score for that motif. This in turn can be very useful to build visual models of overlapping regulation as we presented in **Chapter 1,** Appendix 1 or detection of highly regulated operons as we display in **Chapter 1.**

In **Chapters 4** and **5** I used PCA and t-SNE for dimensional reduction followed by clustering to detect both GO functional groups of interest that showed segregation of species traits such as host or tissue specificity. We used the GODM database which was build using inference and a wisdom of the crowds approach [556] in **Chapters 4** and **5.** Although such transfer of knowledge carries the risks of some false positives, in general it is better to have some information with a few false positives, opposed to having no information at all. As such I want to emphasize that data driven approaches that use inference and wisdom of the crowds' approaches is very useful for to complement hypothesis driven research. Currently detection of segregation was done manually by analysing dendrograms, PCA and t-SNE plots. The approach we deployed in **Chapter 4** was initially unsupervised, leading to detection of non-zoonotic and non-zoonotic strains in *S. suis*. After this initial discovery, we switched to a supervised learning method where we specifically trained Random Forest models to predict zoonotic and non-zoonotic *S. suis* and *S. agalactiae* strains. In **Chapter 5** we used a completely supervised learning approach since we directly trained models to predict tissue and host specificity of different *Mycoplasma's*.

Our approach in both chapters was manual. Large numbers of heatmaps, phylogenetic trees, PCA plots. t-SNE plots and heatmaps were automatically generated, however their analysis was still purely manual. In theory, automation can be applied to detect and segregate patterns or clusters in networks, PCA & t-SNE plots or hierarchical clustering plots such as dendrograms. Methods to automate quantification of segregation of clusters are available, such as K-means, Spectral, clustering, DBSCAN, Partitional Clustering Algorithm based on Nearest Neighbours Heuristics [552], Local Density with Glowworm Swarm Optimization [553], projection to latent structures discriminant analysis [554], extended Minimum Spanning Tree [555] to name a few. Many algorithms for cluster segregation and analysis are present. However, which algorithm works best depends on the data used as well as the chosen parameters for the algorithm [553]. Metrics such Silhouette score or Rand index combined with hyperparameter searches can be used to benchmark the performance of different clustering algorithms on real and *in silico* data sets [553].

No one solution works best in all situations, hence automatic detection of clusters comes with own set of challenges. Additionally, as we see in **Chapter 5,** segregation of classes can be synergistic, meaning that not always one class is segregated perfectly, sometimes a cluster segregates multiple classes from other classes. Even with these challenges, there are still many arguments to be made for automated cluster segregation. Firstly, recognition of groups over multiple dimensions is easier for algorithms than for humans. This is one of the reasons why t-SNE plots in general outperformed PCA plots in this thesis, since they project these multiple dimensions on 2D space where with PCA plots we mostly limited ourselves to explore the first two to three principal components. Secondly, algorithms can scale better than humans. An algorithm can scan many functional groups of proteins that lead to good segregation of many phenotypes and only output relevant results to graphs to be presented to a researcher. An alternative approach to automatic detection of interesting GO functional groups to segregate phenotypes would be to use the feature importance of classifiers to identify groups of functional proteins to further explore.

PCA and t-SNE can also be used to reduce dimensionality before proceeding with classification. Such an approach was not used in this study since it makes the biological interpretation of the classifier rather challenging. It could however be an interesting alternative to the iterative feature reduction approach w used in this thesis. Iterative feature reduction hides the importance of alternative features with the same presence absence since they are not present in the final set of most important features. In general, many interesting methodologies can be use that are developed outside of the field of Bioinformatics and Systems Biology. Machine learning methods are increasingly applied in systems biology. The software tools and documentation for machine learning are more mature than most modelling methods such as Genome Scale Metabolic modelling and dynamic modelling methods used in Systems and Synthetic Biology. Hence, future researchers might choose to use more general applicable methodologies such as machine learning models since they generate reasonable results in fraction of the time required to build fully descriptive models such as GEM models and Dynamic models. Nonetheless this author believes in the future of fully descriptive modelling once proper conditions are met, which I will discuss in the section ‘

Breaking the barrier between models and data’.

7.3.3 Systems Biology Modelling

In this thesis I applied two GEM models, one dynamic model as well as three simple Random Forest classifiers. The dynamic model build in this thesis was initially planned to become part of a whole cell model for *M. pneumonia*. Various strengths, weaknesses, bottlenecks, and solutions were identified for the various modelling approaches. I would like to discuss these various bottlenecks here.

7.3.3.1 Tools and model standards

Unexpectedly , the use of GEM's is not as straight forward as one might expect from a modelling field that exist since 1995 [557]. For example, the CPLEX solver turned out to completely crash the Python kernel, once in every couple of thousands of simulations. Although this bug was reported many years ago, it was not fixed years later at the time of writing **Chapter 4**. Due to extreme ungracefulness of the crash, the only solution to run large number of simulations was to switch from the CPLEX solver to Gurobi [479] solver. Simple parallelisation to scale up simulations were found to be difficult due to the design decisions made by the developers of Cobrapy. For example, making a deep copy of a GEM model within a python environment was found not to result in a deep copy as one would expect. Although these kinds of issues are surmountable and work arounds exist, it indicates a somewhat surprising lack of maturity in the GEM modelling field.

Similarly, when working with dynamic models, a lot of practical troubles were encountered. These troubles included a) the need to manually debug SBML models b) incompatibility and data loss when using valid SBML models with visualisation tools such as Cell Designer [558] and Vanted [559], c) incompatibilities between dynamic models stored as SBML models and tabular conversion formats using SBtab [479]. The above-mentioned incompatibilities force the user to do manual validation and modification of the SBML models each time a model is converted, annotated, or visualised. The holy grail of modelling is often described as having model-driven experimentation. Preferably, multiple iterations of results are there to each time suggest new experiments. Such iteratively approaches are too time consuming to achieve due to the above-mentioned issues unless the model accepts the loss of information at each modelling iteration. Tools such as Copasi PyCoTools [560] and Tellurium [490], [561] are a great step forwards since they enable upscaling and parallelization of dynamic model training, sensitivity analysis and simulations. These software tools are however still in in development with varying build in solvers and as such varying results per version of the software used. Lack of persistent funding still plagues the field of Systems Biology, greatly hampering modelling efforts. Especially modelling efforts that require integration of models, data, and visualisation appears to be not well developed in the experience of this author. At the time of writing, this author does not know a single good solution have full integration of data, models, visualisation. Lack of continuous funding for modelling software results in most modelling tools out there being obsolete, hard to use, incompatible with new data and programming standards. Therefore, making model building and simulations for dynamic models reproducible is near impossible. Although the situation is less dire than a couple of years ago, lack of continuity in tools and standards is still a great threat to future of Systems Biology and modelling in general. Standardization and funding for tools, software and data reuse across difference scientific disciplines as part of the European Open Science Cloud (EOSC) initiative is a good example of work towards good reusability and continued development of software and tools [562].

7.3.3.2 FAIR data management

Having a lot of data, or a great variety of data, does not automatically enable great science. For one, it is hard to link and integrate various data types. Variations in the granularity of data, difference in annotation of the data as well as variations in experimental setup, biological noise, and instrumental noise, can hinder the user of data in Systems Biology Approaches. One can speak to any researcher in the field of Systems Biology to hear their woes on data related problems such as lacking data, lacking metadata, lacking annotation or poor data formats. In fact, studies have shown most researchers use 70-80% of their time on mundane tasks such as finding, accessing and formatting data for reuse [563]. Great science is only possible when using great data. In this author's opinion the lack of Findability, Accessibility, Interoperability and Reusability, are an unacceptable waste of public research funds and a missed opportunity from a scientific point of view since so much data ends up being used only once.

Therefore, in this thesis, a lot of time was spent to make our modelling efforts in **Chapter 6** Findable, Accessible, Interoperable and Reusable (FAIR). I made as much the data and model FAIR by annotating, structuring data and storing it in the FAIRDOM[479] hub and models in the SEEK[478] and BioModels database [564]. Uploading and annotation of most was done long after the generation of this data and not by data generators themselves. I refer to this as "FAIRified data", opposed to truly FAIR data. The data is to "some extend FAIR", meaning mostly there is a focus on Findability and Accessibility while Interoperability and Reusability are still rather lacking due to sparse metadata and limited linking to protocols used to generate the data. The difference between FAIR by design [70] and FAIRified data is visualised in **Figure 29**. While FAIR by design data is a bottom-up approach, FAIRified data is a top-down approach. The challenges encountered, and the time spend on making our dynamic model and its linked data FAIRified, illustrate a) the importance of generating data with standardized annotation and metadata, b) the importance of using automated experimental setups and c) that it is much more efficient for data generators themselves to make their data FAIR. In this author's opinion, FAIR data is a prerequisite for large modelling approaches such as dynamic models of metabolism and whole cell modelling. Currently there is not incentive for data generators and modelers alike to make data and models FAIR. As such, spending time on such a task might be considered 'a wasted of time' since it will not help any researcher to get more publications or finish their PhD in time. However, long term scientific goals can be achieved through standardization and FAIR data and model management. FAIR data management through standardization and large-scale automated data generation is the only way to achieve truly Interoperable and Reusable data and models. The lack of incentive for individual researchers, however, remains a problem that can only be overcome by using automation in experiments and model generation, since that would save time for modelers and researchers.

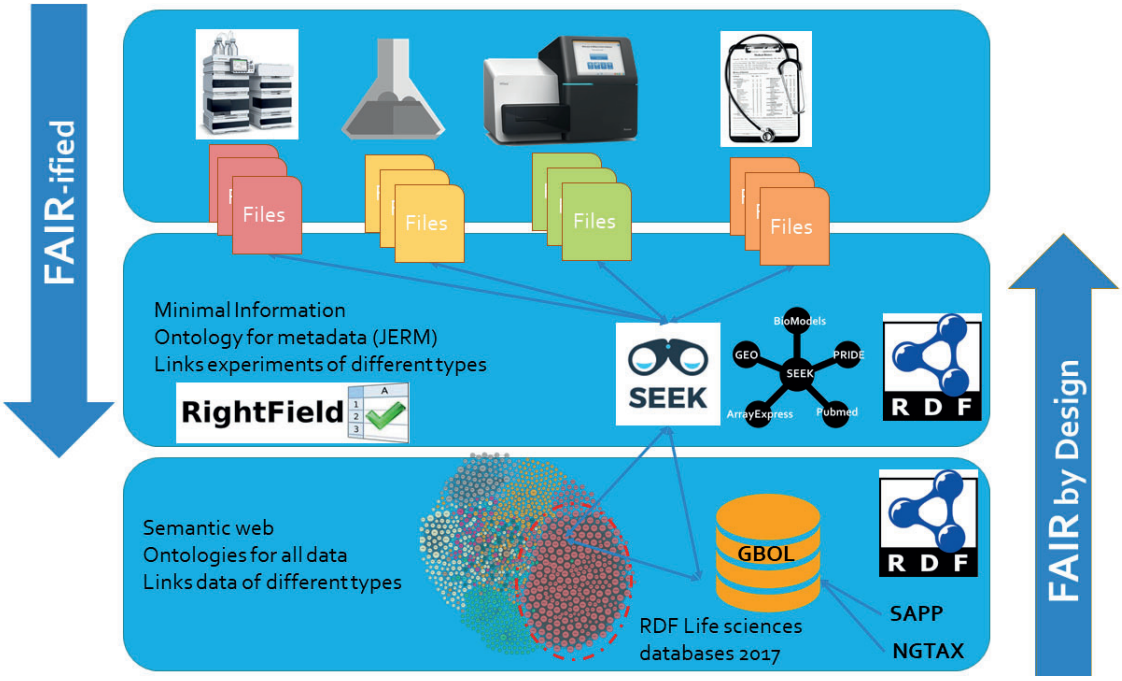

Figure 29. The difference between FAIR by design and FAIRified data.

7.3.3.3 Annotation and metadata

Another problem encountered in modelling is the use of inferior metabolite identifiers [565] as well as lacking annotation in general. In **Chapters 4** and **5**, GEM models were used to map model simulations to determine gene and reaction essentiality with experimentally determined essentiality based on annotation. In both cases incomplete mapping of annotation was observed even though the used models were relatively new and well annotated. Semantic mapping and tools for metabolite mapping are essential to alleviate these issues. Using highly standardized annotation such as domain-based annotation in addition to normal genome annotation might improve the interoperability of models and data. Many improvements have been made in annotation of models in recent years. Examples of improvements made in this field are metabolite mapping tools such as MetaNetX [566], databases such as ChemSpider [567], PubChem [568], Wikidata [569] and automated GEM generation tools with proper annotation such as PathwayTools [54]. There are many types of metabolite identifier a modeller could chose to include in their model. SMILEY and InChI identifiers are however the only identifiers that are uniquely associated to a molecule. SMILEY identifiers have the additional benefits of a) directly capturing the 3d structure of chemical compounds in a human readable format, b) being parsable using regular expressions to identify similar molecules or molecular building groups c) being hierarchical in annotation, allowing models and data with different levels of detail such as isomer specific and non-isomer specific annotation to be mapped with ease. This hierarchy is captured and can be queried in databases such as WikiData [569] and ChEBI [570] which use semantic storage and ontologies to organize the relationships between chemicals. Let us discuss a few examples of why the properties of SMILEY identifiers and the hierarchical storage of metabolite identifiers is so important. In a model instead of using only ATP in a reaction, it could be that other three phosphate

nucleotides such as ATP, CTP, GTP and UTP are used. Representing such a reaction is relatively simple when using SMILEY identifiers and regular expressions in rule based and agent-based modelling. Another example would be to query enzyme Monod constants for similar metabolites using SMILEY identifiers and regular expression searches. An example would be the binding of small molecules with phosphate groups, which is mostly dominated by the nature of the phosphate group and the resonance stability it provides to molecules. Using SMILEY identifiers and simple regular expressions, a database can be queries on km values of other small molecule with similar side groups. Many more examples can be thought of. By providing these examples I hope to illustrate why choosing the right identifiers for both data and models can be instrumental to Systems Biology which relies heavily on the integration of data and models. In this authors opinion, using 'gold level metabolite identifiers' such as SMILEY, InChI, InChIKey's or ChEBI identifiers, should become an obligatory requirement for funding of research that involves measurements or modelling of metabolites. Complex molecules such as DNA, RNA and other biopolymers remain challenging to annotate in a standardized way. Regular expressions based on the SMILEY or InChIKey identifiers which are synonymous with the structure of the molecules they represent could provide a solution here. Examples of such an approach are ClassyFire, a computer program for automated chemical classification in the using approximately 200 regular expression searches and rules in order to detect structural features [553]. For their classification they use the SMART molecular pattern matching language which is related to SMILES molecular language and chemical classification is stored in the ChemOnt ontology. A regular expression for glycogen could match branches of different lengths which consists of Glucose residues linked linearly by α-1,4 glycosidic bonds, as well as branches of different lengths which branch off via α-1,6 glycosidic linkages. Regular expressions can allow for a range of lengths in the chain while restricting to long chains to separate glycogen from starch which is structurally similar, but which has fewer branches and is less compact than glycogen. Capturing all possible configuration of complex biomolecules as individual structures might become data intensive, while a pattern that generalizes these structures takes only a single line. RetroPath 2.0 is an example of a tool that uses structure based SMART patterns to identify generalized metabolite reactions and pathways to aid reaction and pathway mapping and metabolic engineering efforts that exploit enzyme promiscuity [571]. Differences and similarities can be captured in a SMILEY, SMART or InChiKey based patterns. Automated methods such as ClassyFire structure molecules in a hierarchical ontology based on patterns in their stucture. An added benefit of using regular expressions based on the chemical structure, would be that these patterns are both machine and human readable and can for example be used in agent-based models with reactions that work with multiple alternative metabolites in a single reaction. A single reaction can unambiguously represent a whole range of possible reactions in a compressed and easy to interpret way. If different rates for different forms of similar molecules is needed, the modeler can break up the large regular expressions into smaller more restrictive regular expressions or even chemical identifiers to specify the rates for each variant of the molecule. An example in modelling would be the use of metabolite patterns to specify the use of different mono- di- or three-phosphate nucleotides as well as large biomolecules with different possible structures such as the glycogen example given before. This is also possible using the

parent child relationship in chemical databases that use hierarchical ontologies to retrieve all 'descendants' in a tree of metabolites. However, regular expressions have the benefits of being programming agnostic and not requiring interaction with any database or being specific for any ontology. A last example of where generalised metabolite patterns might be useful, is for linking and unifying different chemical ontologies, databases, and identifiers. Structure based patterns can be used to compare branches of related chemical compounds independent of the structure of a database and can for example identity metabolites present in one database that can be included in another database based on fitting a generalised pattern that matches the ontology. In summary, using structure-based patterns with SMILEY identifiers, SMART identifiers or InchiKey identifiers, can be used to unambiguously classifying chemicals as was demonstrated in ClassyFire and RetroPath 2.0 and can be used for metabolite, reaction, pathway, model, database, and ontology mapping.

7.3.3.4 Robotics to automate, standardize and upscale experimentation

As mentioned in the previous section, there is little incentive for individual researchers to manually annotate data and models to be FAIR. Improvements in standardization of annotation of data and models can come most easily from automation since most models are at least partly automatically generated before gap filling and manual curation by a modeler. Recent years have shown an increase in projects with standardized and large-scale data generation. The use of large-scale automated experiments does not only result in better annotated and FAIR data, but also results in data with much lower amounts of technical noise. Additionally, when using automation to generate multiple omics data from a single experiment, sampled at a single time point, biological noise can also be greatly reduced. Automation and large-scale data generation might therefore be key to truly make FAIR data management a standard. Finally, systems biologist might have the quality and quantity in data to make models live up to their promises. The use of best practices in annotation, metadata, and modelling formats could make it even easier to build pipelines and modelling tools that interact with them. Standardization greatly simplifies the work of linking data and models while greatly reducing the time for scientist to go from a biological question to biological findings and publications. In the opinion of this author, standardization in data and model generation is the only way to efficiently make the iterative modelling cycle work for large quantitative models.

7.3.4 Breaking the barrier between models and data

Systems biology is an integrative study field. Therefore, research stands and falls with a) the availability and quality of data and b) the availability and quality of annotation and metadata and c) the ease of interoperability of data and models. As I pointed out previously, interoperability and reusability rely on standardization in experiments, data, models, annotation, and tools. From the point of view of this author there should be no separation between models and data. Genomic data is used to build GEM models, omics data can be used to generate and impose constraints on models or provide insight by plotting them on 'ordered networks' as discussed in the sub-chapter 'Network and clustering. The results of models can be compared to experimental data such as essentiality data or be visualised on ordered maps such as pathway maps. All

these mentioned operations require interoperability and as such standardization in annotation. When fully embracing an engineering approach one should take the perspective of programmers and software engineers. From a software engineering principal molecules and reactions are simply two types of data objects which can link to various properties and identifiers. Discrete models and pathway maps use instantiation with multiple such objects existing at the same time. For example, a map can show ATP at multiple locations without showing the link to all other instances of ATP in a reaction map, unless the user choses to see them. Similarly, agent-based models throw the dice for each instance of a molecule involved in a reaction. Continuous models on the other hand use single instantiations of objects while pathway maps allow multiple instances of an object to exist at the same time. In their essence, all these different modelling types still exist out of two types of objects, metabolites, and reactions which can have many properties such as identifiers or reactions modelled at different levels of abstraction. From an engineering or software engineering perspective, these properties are extremely simple and can be captured with ease in a single data/object structure with a simple ontology. Petri Nets are an early effort that shows it can be easy to combine different modelling types by embracing the graph nature of models of being a) a physical object such as a molecule or protein or b) being an action/abstract object such as a reaction [572], [573]. Lessons can be learned from this early attempt such as embracing a most minimal ontology with extendible properties such as annotation and graphical representation. Being simple, minimal, and extensible are key properties for any data and modelling standard. The flexibility of using minimal ontologies is exemplified by PetriNets as they are used to model many type of systems, biological [574]–[576], automatic Web service composition using fuzzy logic [577], [578] as well as the fact that PetriNets themselves were early adopters of semantic web standards and ontologies as is the case in OPENET[579] or the Petri Net Ontology [572].

Due to the graph nature of models, RDF and semantic data storage would be a logical choice. Models should not be partly semantic data (BioPAX), or after creation converted to a hard to read, hard to use and hard to visualize container format such as SBML. Instead, models should ab-initio be stored in well-defined interoperable semantic data structures. Switching to semantic data standards as a default can greatly decrease the woes of incompatibilities encountered in modelling. The advantage of Semantically storing of model data is that it enables much better machine interoperability, data validation and downstream development. Data quality of such models can be checked by using code generators such as EMPUSA which enforce proper heredity and data types on RDF data similar to how this is enforced in Object oriented programming [389]. Semantically stored models can be converted into executable models at any time with very little computation effort since a single model account only for a very small number of triples and querying RDF is highly optimised. Through their graph nature, models stored as semantic data are more scalable and easier to interact with by both users and software alike. Querying or simulating a thousand models or a single model would roughly cost the same amount of user effort and would approximately scale linearly [580]. Human time is the most limited resource; hence we should fully use all technological tricks we can to simply data analysis and modelling by removing all boundaries between data and models by

making them both human and machine findable and interoperable by using semantic data storage and ontologies.

But hey, would switching to RDF not create yet another modelling standard? This author would argue not. Basically, because I am not defining a new modelling standard, but I bring down existing modelling and data standards to their minimum and capturing them in the most minimal data structure possible, namely RDF which is basically *object-property-object*. Most successful semantic data ontologies are very simple in their nature while being fully extendible with link out properties to other database and other ontologies. While semantically stored models are easy to convert to SBML, this conversion would probably be loss full. Hence, storing all models as RDF objects would be desirable since SBML is more limited and has no advantage over models stored as RDF.

Unifying different modelling types is in a precondition for complex modelling approaches such as whole cell modelling. The lack of a good standard and formats to support whole cell modelling and the lack of proper tools to build and validate these models are mentioned as major bottlenecks as can be seen by a survey among hundreds of modellers [581]. A lot of progress has been made over the years that are in line with the data-centric philosophy argued for by this author. RDF conversion of models from the BioModels database to BioPax was used to query multiple models visualize the occurrence of reactions over multiple models on a pathway maps and to cluster models based on their similarity [582]–[586].

Automatic generation of models with tools such as Pathway Tools that uses MetaCyc annotation [54] or building upon models with standard annotation as found in BiGG Models [56]that standardized identifiers called BiGG IDs and SEEK [55] model database where Tools are available for model annotation in accordance with the MIRIAM guidelines for annotation. Such tools and databases have become the starting point for most modelers. Semantic annotation stored in Research Description Framework BIOPAX has become the de facto standard for model annotation, greatly improving the interoperability and reusability of models [587]. Models within Seek can be queried via their metadata as part of SysMO-SEEK metadata, however, the reactions within models themselves cannot yet be queried. To extend these developments and to switch to a complete semantic framework to store omics data, model data and pathway maps, is therefore sensible and in line with developments in recent years. Switching to semantically stored models should allow for a more efficient workflow when building, querying, and interacting with these models. For example, the RDF JSON format is much more readable and easier to interact with programmatically than SBML.

In my opinion the combination of a) moving to completely semantic storage of models, b) switching to unambiguous human and machine readable identifiers combined with an ontology to capture the hierarchy in annotation such as SMILEY identifiers, InChI and ChEBI identifiers, hashed identifiers based on the structure such as InChIKey's c) development of scalable modelling tools that allow parallelization such as libroadrunner [491], Tellurium[490] and PycoTools [560], automated model building, validation, visualization and d) the use of large-scale automated data generation by bio-foundries, could lead to a revival of the Systems Biology modelling field and would

enable the conditions for larger and more complex models such as models of the whole cell [588]. Using the above best practices in annotation and FAIR data management will enable easier data integration, data exploration and better modelling of pathogen host interaction. It is my believe that only using such best practice can we hope to answer the question *What are the patterns in bacterial pathogen host interaction?"*. I believe that enforcing the above suggested best practices, larger and more complex modelling approaches such as Whole Cell modelling will become attainable.

Summary

This thesis aims to answer the question *"What are the patterns in bacterial pathogen host interaction?"*. Various pathogens are studied using common concepts, methods, and strategies. In this thesis we studied the pathogenesis of *Mycobacterium tuberculosis, Mycoplasma pneumoniae, Staphylococcus* and *Streptococcus* species such as *Staphylococcus aureus* and *Streptococcus pneumoniae*.

Chapter 1 discusses the bacterial pathogenesis and the pathogens studied within this thesis. The strategies deployed by these bacterial pathogens are analysed as well as their impact on society. Zoonotic pathogens, which can infect both humans and animals, are discussed. There is an abundance of data available on the bacteria studied in this thesis. This abundance in data makes them particularly suitable for Systems Biology approaches. Systems Biology involves integration of multiple 'omics' data such as genomics, transcriptomics, proteomics, metabolomics and phenomics as well as modelling. This chapter discusses the differences between reductionistic approaches, which study biological parts, and holistic approaches (e.g., Systems Biology), which study systems at large, such as networks, and the properties that emerge from the interactions in a system. Within this thesis I use a Systems Biology approach to study large systems and identify emerging properties. At the same time, I use reductionistic approach by emphasizing modularity in biological systems, and by imposing structure on the biology by creating maps. This thesis balances the use of these two different philosophical approaches since biological systems are chaotic and leaky as well as ordered and modular. Common Systems Biology methodologies are discussed, as well as the importance of FAIR data management on scientific research and long term societal impact of scientific data.

Chapter 2 identifies three major virulence strategies in *M. tuberculosis*. *M. tuberculosis* was responsible for an approximate 1.5 million deaths in non-HIV patients in 2021. To better understand *M. tuberculosis* and to find new drug and vaccine KO candidates, literature and omics data were analysed using synchronous network visualization. The output of this chapter is a visual and modular overview of the three virulence strategies, their components, and their regulation. Moreover, this chapter identifies a single regulatory cascade for these three virulence strategies that respond to environmental cues like limited availability of divalent metals in the phagosome.

Chapter 3 provides three examples of how Meme2Fimo and Synchronous Network Data Integration (SyNDI) framework, were used to detect multiple and related binding motifs within the DevR regulon, which have not yet been described in the literature, including a motif that is related to the *M. tuberculosis* regulator SigE.

Chapter 4 compares 235 *Staphylococcus* and 315 *Streptococcus* genomes based on their protein domain content. This chapter shows the relationships between protein persistence and essentiality by integrating essentiality predictions from two metabolic models and essentiality measurements from six large-scale transposon mutagenesis experiments. Clusters of strains within species were identified based on proteins

associated to similar biological processes. These two different clusters correspond to zoonotic and non-zoonotic strains. Two Random Forest classifiers were built that predicted the zoonotic potential of *Streptococcus suis* and *Streptococcus agalactiae*. Furthermore, this chapter identifies shared attributes between of *Staphylococcus aureus* and *Streptococcus pyogenes* that allow them to cause necrotizing fasciitis.

Chapter 5 investigates which *Mycoplasma* proteins are most predictive of tissue and host trophism and to which functional groups of proteins they belong. I retrieved and annotated 432 *Mycoplasma* genomes and combined their genome information with host and tissue isolation data. I compared clustering of *Mycoplasma* and *M. pneumoniae* strains based on different biological process functional groups of proteins. This chapter shows that proteins belonging to the Gene Ontology (GO) Biological process group '*Interspecies interaction between organisms*' proteins are most important for predicting the pathogenesis of *Mycoplasma* strains, while those belonging to *'Quorum sensing'* and *'Biofilm formation'* proteins are most important for predicting pathogenesis of *M. pneumoniae*. Two Random Forest Classifiers were trained to accurately predicts host and tissue specificity based on only 12 proteins. For *Mycoplasma* host specificity CTP synthase complex, magnesium transporter MgtE, and glycine cleavage system are most important for correctly classifying *Mycoplasma* strains that infect humans including opportunistic zoonotic strains. For tissue specificity, this chapter found that a) known virulence and adhesions factor Methionine sulphate reductase MetA is predictive of urinary tract *infecting Mycoplasmas,* b) an extra cytoplasmic thiamine binding lipoprotein is most predictive of gastro-intestinal infecting *Mycoplasmas,* c) a type I restriction endonuclease is most predictive of respiratory infecting *Mycoplasmas,* and d) a branched-chain amino acid transport system is most predictive for blood infecting *Mycoplasmas*.

Chapter 6 explores the adaptability and robustness of glycolysis and pyruvate metabolism of *Mycoplasma pneumoniae* (MPN). Dual approaches were used in this chapter. Firstly, this chapter analysed metabolomics data collected for many OE and KO mutants and perturbation samples. Secondly, this chapter trained a dynamic model of central carbon metabolism and tested the model's capacity to predict these mutants and perturbation samples as well as identify key controlling factors in central carbon metabolism. The analysis of metabolite data as well as the dynamic model analysis indicate MPN metabolism is inherently robust against perturbations due to its network structure. Two key control hubs of central carbon metabolism were identified.

Chapter 7 discusses the results of the various chapters and how they answer the research question of this thesis. The similarities and differences in strategies of the studied bacterial pathogens are discussed. This chapters concludes that although pathogenesis might vary between bacterial pathogens, common strategies and building blocks are used, leading to patterns in pathogenesis. These patterns are to some extend transferable between bacteria and can be used to elucidate bacterial host interaction.

References

[1] S.-W. Jeong and Y. J. Choi, "Extremophilic Microorganisms for the Treatment of Toxic Pollutants in the Environment," *Molecules*, vol. 25, no. 21, p. 4916, Oct. 2020, doi: 10.3390/molecules25214916.

[2] N. T. Enyedi *et al.*, "Radioactive environment adapted bacterial communities constituting the biofilms of hydrothermal spring caves (Budapest, Hungary)," *J. Environ. Radioact.*, vol. 203, no. September 2018, pp. 8–17, Jul. 2019, doi: 10.1016/j.jenvrad.2019.02.010.

[3] R. Khan, F. C. Petersen, and S. Shekhar, "Commensal Bacteria: An Emerging Player in Defense Against Respiratory Pathogens," *Front. Immunol.*, vol. 10, no. MAY, pp. 1–9, May 2019, doi: 10.3389/fimmu.2019.01203.

[4] J. M. Kwiecinski and A. R. Horswill, "Staphylococcus aureus bloodstream infections: pathogenesis and regulatory mechanisms," *Curr. Opin. Microbiol.*, vol. 53, pp. 51–60, Feb. 2020, doi: 10.1016/j.mib.2020.02.005.

[5] A. Reiss-Mandel and G. Regev-Yochay, "Staphylococcus aureus and Streptococcus pneumoniae interaction and response to pneumococcal vaccination: Myth or reality?," *Hum. Vaccin. Immunother.*, vol. 12, no. 2, pp. 351–357, Feb. 2016, doi: 10.1080/21645515.2015.1081321.

[6] T. L. Keiser and G. E. Purdy, "Killing Mycobacterium tuberculosis In Vitro : What Model Systems Can Teach Us," *Microbiol. Spectr.*, vol. 5, no. 3, pp. 139–148, May 2017, doi: 10.1128/microbiolspec.TBTB2-0028-2016.

[7] J. a. H. Wodke *et al.*, "MyMpn: a database for the systems biology model organism Mycoplasma pneumoniae," *Nucleic Acids Res.*, vol. 43, no. D1, pp. D618–D623, Jan. 2015, doi: 10.1093/nar/gku1105.

[8] WHO, *Global tuberculosis report 2021, WHO*. [Online]. Available: https://www.who.int/publications/i/item/9789240037021

[9] M. Gengenbacher and S. H. E. Kaufmann, "Mycobacterium tuberculosis : success through dormancy," *FEMS Microbiol. Rev.*, vol. 36, no. 3, pp. 514–532, May 2012, doi: 10.1111/j.1574-6976.2012.00331.x.

[10] S. Sharma and J. S. Tyagi, "Mycobacterium tuberculosis DevR/DosR Dormancy Regulator Activation Mechanism: Dispensability of Phosphorylation, Cooperativity and Essentiality of α10 Helix," *PLoS One*, vol. 11, no. 8, p. e0160723, Aug. 2016, doi: 10.1371/journal.pone.0160723.

[11] N. Zondervan, J. van Dam, P. Schaap, V. Martins dos Santos, and M. Suarez-Diez, "Regulation of Three Virulence Strategies of Mycobacterium tuberculosis: A Success Story," *Int. J. Mol. Sci.*, vol. 19, no. 2, p. 347, Jan. 2018, doi: 10.3390/ijms19020347.

[12] Centers for Disease Control and Prevention, "Deadly Staph Infections Still Threaten the U.S. | CDC Online Newsroom | CDC." https://www.cdc.gov/media/releases/2019/p0305-deadly-staph-infections.html (accessed Jun. 18, 2022).

[13] L. R. K. Brooks and G. I. Mias, "Streptococcus pneumoniae's Virulence and Host Immunity: Aging, Diagnostics, and Prevention," *Front. Immunol.*, vol. 9, no. JUN, Jun. 2018, doi: 10.3389/fimmu.2018.01366.

[14] Z.-R. Lun, Q.-P. Wang, X.-G. Chen, A.-X. Li, and X.-Q. Zhu, "Streptococcus suis: an emerging zoonotic pathogen," *Lancet Infect. Dis.*, vol. 7, no. 3, pp. 201–209, Mar.

2007, doi: 10.1016/S1473-3099(07)70001-4.

[15] A. C. N. Botelho, A. F. M. Ferreira, S. E. L. Fracalanzza, L. M. Teixeira, and T. C. A. Pinto, "A Perspective on the Potential Zoonotic Role of Streptococcus agalactiae: Searching for a Missing Link in Alternative Transmission Routes," *Front. Microbiol.*, vol. 9, no. March, pp. 1–5, Mar. 2018, doi: 10.3389/fmicb.2018.00608.

[16] J. He *et al.*, "Insights into the pathogenesis of Mycoplasma pneumoniae," *Mol. Med. Rep.*, vol. 14, no. 5, pp. 4030–4036, Nov. 2016, doi: 10.3892/mmr.2016.5765.

[17] R. Rosengarten *et al.*, "Host-pathogen interactions in mycoplasma pathogenesis: Virulence and survival strategies of minimalist prokaryotes," *Int. J. Med. Microbiol.*, vol. 290, no. 1, pp. 15–25, Mar. 2000, doi: 10.1016/S1438-4221(00)80099-5.

[18] S. Razin and L. Hayflick, "Highlights of mycoplasma research—An historical perspective," *Biologicals*, vol. 38, no. 2, pp. 183–190, Mar. 2010, doi: 10.1016/j.biologicals.2009.11.008.

[19] B. B. A. Raymond *et al.*, "Mycoplasma hyopneumoniae resides intracellularly within porcine epithelial cells," *Sci. Rep.*, vol. 8, no. 1, p. 17697, Dec. 2018, doi: 10.1038/s41598-018-36054-3.

[20] J. a H. Wodke *et al.*, "Dissecting the energy metabolism in Mycoplasma pneumoniae through genome-scale metabolic modeling," *Mol. Syst. Biol.*, vol. 9, no. 1, p. 653, Jan. 2013, doi: 10.1038/msb.2013.6.

[21] E. Gaspari *et al.*, "Model-driven design allows growth of Mycoplasma pneumoniae on serum-free media," *npj Syst. Biol. Appl.*, vol. 6, no. 1, p. 33, Dec. 2020, doi: 10.1038/s41540-020-00153-7.

[22] L. Sun, Tzu, Giles, *The Art of War*. Filiquarian Publishing, LLC., 2012. [Online]. Available: https://books.google.nl/books?id=1bApuAAACAAJ

[23] P. M. Sharp and B. H. Hahn, "Origins of HIV and the AIDS Pandemic," *Cold Spring Harb. Perspect. Med.*, vol. 1, no. 1, pp. a006841–a006841, Sep. 2011, doi: 10.1101/cshperspect.a006841.

[24] A. Pizam, "The aftermath of the corona virus pandemic," *Int. J. Hosp. Manag.*, vol. 95, no. 95,102909, p. 102909, May 2021, doi: 10.1016/j.ijhm.2021.102909.

[25] J. S. Mackenzie and D. W. Smith, "COVID-19: a novel zoonotic disease caused by a coronavirus from China: what we know and what we don't," *Microbiol. Aust.*, vol. 41, no. 1, p. 45, 2020, doi: 10.1071/MA20013.

[26] S. A. Khan, M. A. Imtiaz, M. M. Islam, A. Z. Tanzin, A. Islam, and M. M. Hassan, "Major bat-borne zoonotic viral epidemics in Asia and Africa: A systematic review and meta-analysis," *Vet. Med. Sci.*, pp. 1–15, May 2022, doi: 10.1002/vms3.835.

[27] J. Cohen, "Monkeypox outbreak questions intensify as cases soar," *Science (80-.).*, vol. 376, no. 6596, pp. 902–903, 2022, doi: 10.1126/SCIENCE.ADD1068.

[28] C. Chakraborty, M. Bhattacharya, S. S. Nandi, R. K. Mohapatra, K. Dhama, and G. Agoramoorthy, "Appearance and re-appearance of zoonotic disease during the pandemic period: long-term monitoring and analysis of zoonosis is crucial to confirm the animal origin of SARS-CoV-2 and monkeypox virus," *Vet. Q.*, vol. 42, no. 1, pp. 119–124, Jun. 2022, doi: 10.1080/01652176.2022.2086718.

[29] Y. Ke, Z. Chen, and R. Yang, "Yersinia pestis: mechanisms of entry into and resistance to the host cell," *Front. Cell. Infect. Microbiol.*, vol. 3, no. DEC, pp. 1–9, 2013, doi: 10.3389/fcimb.2013.00106.

[30] P. C. F. Oyston and K. E. Isherwood, "The many and varied niches occupied by Yersinia pestis as an arthropod-vectored zoonotic pathogen," *Antonie Van Leeuwenhoek*, vol. 87, no. 3, pp. 171–177, Apr. 2005, doi: 10.1007/s10482-004-4619-3.

[31] I. Morozova *et al.*, "New ancient Eastern European Yersinia pestis genomes illuminate the dispersal of plague in Europe," *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 375, no. 1812, p. 20190569, Nov. 2020, doi: 10.1098/rstb.2019.0569.

[32] L. Christou, "The global burden of bacterial and viral zoonotic infections," *Clin. Microbiol. Infect.*, vol. 17, no. 3, pp. 326–330, Mar. 2011, doi: 10.1111/j.1469-0691.2010.03441.x.

[33] A. Aderem, "Systems Biology: Its Practice and Challenges," *Cell*, vol. 121, no. 4, pp. 511–513, May 2005, doi: 10.1016/j.cell.2005.04.020.

[34] L. Hood and L. Rowen, "The human genome project: big science transforms biology and medicine," *Genome Med.*, vol. 5, no. 9, p. 79, 2013, doi: 10.1186/gm483.

[35] B. C. Daniels, Y.-J. Chen, J. P. Sethna, R. N. Gutenkunst, and C. R. Myers, "Sloppiness, robustness, and evolvability in systems biology," *Curr. Opin. Biotechnol.*, vol. 19, no. 4, pp. 389–395, Aug. 2008, doi: 10.1016/j.copbio.2008.06.008.

[36] W. B. Copeland *et al.*, "Computational tools for metabolic engineering," *Metab. Eng.*, vol. 14, no. 3, pp. 270–280, May 2012, doi: 10.1016/j.ymben.2012.03.001.

[37] A. Trewavas, "A Brief History of Systems Biology," *Plant Cell*, vol. 18, no. 10, pp. 2420–2430, Oct. 2006, doi: 10.1105/tpc.106.042267.

[38] D. Müller, L. Aguilera-Vázquez, M. Reuss, and K. Mauch, "Integration of metabolic and signaling networks," in *Systems Biology*, vol. 13, Berlin/Heidelberg: Springer-Verlag, 2007, pp. 235–256. doi: 10.1007/b136529.

[39] A. Kremling and J. Saez-Rodriguez, "Systems biology—An engineering perspective," *J. Biotechnol.*, vol. 129, no. 2, pp. 329–351, Apr. 2007, doi: 10.1016/j.jbiotec.2007.02.009.

[40] R. Breitling, D. Gilbert, M. Heiner, and R. Orton, "A structured approach for the engineering of biochemical network models, illustrated for signalling pathways," *Brief. Bioinform.*, vol. 9, no. 5, pp. 404–421, Apr. 2008, doi: 10.1093/bib/bbn026.

[41] M. D. Wilkinson *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship," *Sci. Data*, vol. 3, no. 1, p. 160018, Dec. 2016, doi: 10.1038/sdata.2016.18.

[42] R. D. Finn *et al.*, "The Pfam protein families database: towards a more sustainable future," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D279–D285, Jan. 2016, doi: 10.1093/nar/gkv1344.

[43] T. L. Bailey, J. Johnson, C. E. Grant, and W. S. Noble, "The MEME Suite," *Nucleic Acids Res.*, vol. 43, no. W1, pp. W39–W49, Jul. 2015, doi: 10.1093/nar/gkv416.

[44] N. A. Zondervan, V. A. P. Martins dos Santos, M. Suarez-Diez, and E. Saccenti, "Phenotype and multi-omics comparison of Staphylococcus and Streptococcus uncovers pathogenic traits and predicts zoonotic potential," *BMC Genomics*, vol. 22, no. 1, p. 102, Dec. 2021, doi: 10.1186/s12864-021-07388-6.

[45] J. E. Galagan *et al.*, "The Mycobacterium tuberculosis regulatory network and hypoxia," *Nature*, vol. 499, no. 7457, pp. 178–183, Jul. 2013, doi: 10.1038/nature12337.

[46] H. N. B. Moseley, “ERROR ANALYSIS AND PROPAGATION IN METABOLOMICS DATA ANALYSIS,” *Comput. Struct. Biotechnol. J.*, vol. 4, no. 5, p. e201301006, Jan. 2013, doi: 10.5936/csbj.201301006.

[47] J. J. Faith *et al.*, “Large-Scale Mapping and Validation of Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles,” *PLoS Biol.*, vol. 5, no. 1, p. e8, Jan. 2007, doi: 10.1371/journal.pbio.0050008.

[48] M. Suarez-Diez and E. Saccenti, “Effects of Sample Size and Dimensionality on the Performance of Four Algorithms for Inference of Association Networks in Metabonomics,” *J. Proteome Res.*, vol. 14, no. 12, pp. 5119–5130, Dec. 2015, doi: 10.1021/acs.jproteome.5b00344.

[49] P. K. Robinson, “Enzymes: principles and biotechnological applications,” *Essays Biochem.*, vol. 59, pp. 1–41, Nov. 2015, doi: 10.1042/bse0590001.

[50] S. Jahagirdar and E. Saccenti, “Evaluation of Single Sample Network Inference Methods for Metabolomics-Based Systems Medicine,” *J. Proteome Res.*, vol. 20, no. 1, pp. 932–949, Jan. 2021, doi: 10.1021/acs.jproteome.0c00696.

[51] E. Lindfors, J. C. J. van Dam, C. M. C. Lam, N. A. Zondervan, V. A. P. Martins dos Santos, and M. Suarez-Diez, “SyNDI: synchronous network data integration framework,” *BMC Bioinformatics*, vol. 19, no. 1, p. 403, Dec. 2018, doi: 10.1186/s12859-018-2426-5.

[52] J. C. van Dam, P. J. Schaap, V. A. Martins dos Santos, and M. Suárez-Diez, “Integration of heterogeneous molecular networks to unravel gene-regulation in Mycobacterium tuberculosis,” *BMC Syst. Biol.*, vol. 8, no. 1, p. 111, Dec. 2014, doi: 10.1186/s12918-014-0111-5.

[53] G. Michal and D. Schomburg, *Biochemical pathways: an atlas of biochemistry and molecular biology*, 2end ed. Wiley, 2013.

[54] P. D. Karp *et al.*, “Pathway Tools version 23.0 update: software for pathway/genome informatics and systems biology,” *Brief. Bioinform.*, vol. 22, no. 1, pp. 109–126, Jan. 2021, doi: 10.1093/bib/bbz104.

[55] K. Wolstencroft *et al.*, “SEEK: a systems biology data and model management platform,” *BMC Syst. Biol.*, vol. 9, no. 1, p. 33, Dec. 2015, doi: 10.1186/s12918-015-0174-y.

[56] Z. A. King *et al.*, “BiGG Models: A platform for integrating, standardizing and sharing genome-scale models,” *Nucleic Acids Res.*, vol. 44, no. D1, pp. D515–D522, Jan. 2016, doi: 10.1093/nar/gkv1049.

[57] K. Sriyudthsak, F. Shiraishi, and M. Y. Hirai, “Mathematical Modeling and Dynamic Simulation of Metabolic Reaction Systems Using Metabolome Time Series Data,” *Front. Mol. Biosci.*, vol. 3, no. May, p. 15, May 2016, doi: 10.3389/fmolb.2016.00015.

[58] A. Gábor and J. R. Banga, “Robust and efficient parameter estimation in dynamic models of biological systems,” *BMC Syst. Biol.*, vol. 9, no. 1, p. 74, Dec. 2015, doi: 10.1186/s12918-015-0219-2.

[59] J. D. Orth, I. Thiele, and B. Ø. Palsson, “What is flux balance analysis?,” *Nat. Biotechnol.*, vol. 28, no. 3, pp. 245–248, Mar. 2010, doi: 10.1038/nbt.1614.

[60] Y. Kuriya and M. Araki, “Dynamic flux balance analysis to evaluate the strain production performance on shikimic acid production in Escherichia coli,” *Metabolites*, vol. 10, no. 5, 2020, doi: 10.3390/metabo10050198.

[61] D. Gilbert, M. Heiner, Y. Jayaweera, and C. Rohr, “Towards dynamic genome-scale

models," *Brief. Bioinform.*, vol. 20, no. 4, pp. 1167–1180, Jul. 2019, doi: 10.1093/bib/bbx096.

[62] W. Poncheewin, G. D. A. Hermes, J. C. J. van Dam, J. J. Koehorst, H. Smidt, and P. J. Schaap, "NG-Tax 2.0: A Semantic Framework for High-Throughput Amplicon Analysis," *Front. Genet.*, vol. 10, no. January, pp. 1–12, Jan. 2020, doi: 10.3389/fgene.2019.01366.

[63] B. Mons, C. Neylon, J. Velterop, M. Dumontier, L. O. B. da Silva Santos, and M. D. Wilkinson, "Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud," *Inf. Serv. Use*, vol. 37, no. 1, pp. 49–56, Mar. 2017, doi: 10.3233/ISU-170824.

[64] M. Bloemers and A. Montesanti, "The FAIR Funding Model: Providing a Framework for Research Funders to Drive the Transition toward FAIR Data Management and Stewardship Practices," *Data Intell.*, vol. 2, no. 1–2, pp. 171–180, Jan. 2020, doi: 10.1162/dint_a_00039.

[65] J. J. Koehorst, J. C. J. van Dam, E. Saccenti, V. A. P. Martins dos Santos, M. Suarez-Diez, and P. J. Schaap, "SAPP: functional genome annotation and analysis through a semantic framework using FAIR principles," *Bioinformatics*, vol. 34, no. 8, pp. 1401–1403, Apr. 2018, doi: 10.1093/bioinformatics/btx767.

[66] M. L. Neal *et al.*, "Harmonizing semantic annotations for computational models in biology," *Brief. Bioinform.*, vol. 20, no. 2, pp. 540–550, Mar. 2019, doi: 10.1093/bib/bby087.

[67] K. Wolstencroft *et al.*, "FAIRDOMHub: a repository and collaboration environment for sharing systems biology research," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D404–D407, Jan. 2017, doi: 10.1093/nar/gkw1032.

[68] J. Wise *et al.*, "Implementation and relevance of FAIR data principles in biopharmaceutical R&D," *Drug Discov. Today*, vol. 24, no. 4, pp. 933–938, Apr. 2019, doi: 10.1016/j.drudis.2019.01.008.

[69] M. Panahiazar, V. Taslimitehrani, A. Jadhav, and J. Pathak, "Empowering personalized medicine with big data and semantic web technology: Promises, challenges, and use cases," in *2014 IEEE International Conference on Big Data (Big Data)*, Oct. 2014, pp. 790–795. doi: 10.1109/BigData.2014.7004307.

[70] R. Kleerebezem, G. Stouten, J. Koehorst, A. Langenhoff, P. Schaap, and H. Smidt, "Experimental infrastructure requirements for quantitative research on microbial communities," *Curr. Opin. Biotechnol.*, vol. 67, pp. 158–165, Feb. 2021, doi: 10.1016/j.copbio.2021.01.017.

[71] D. C. Berrios, A. Beheshti, and S. V. Costes, "FAIRness and Usability for Open-access Omics Data Systems," *AMIA ... Annu. Symp. proceedings. AMIA Symp.*, vol. 2018, pp. 232–241, 2018.

[72] B. McBride, "The Resource Description Framework (RDF) and its Vocabulary Description Language RDFS," in *Handbook on Ontologies*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 51–65. doi: 10.1007/978-3-540-24750-0_3.

[73] C. E. Grant, T. L. Bailey, and W. S. Noble, "FIMO: scanning for occurrences of a given motif," *Bioinformatics*, vol. 27, no. 7, pp. 1017–1018, Apr. 2011, doi: 10.1093/bioinformatics/btr064.

[74] WHO, "Globa Tuberculosis report 2017," 2017.

[75] L. S. Meena and Rajni, "Survival mechanisms of pathogenic Mycobacterium

tuberculosis H37Rv," *FEBS J.*, vol. 277, no. 11, pp. 2416–2427, Jun. 2010, doi: 10.1111/j.1742-4658.2010.07666.x.

[76] W. H. Conrad *et al.*, "Mycobacterial ESX-1 secretion system mediates host cell lysis through bacterium contact-dependent gross membrane disruptions," *Proc. Natl. Acad. Sci.*, vol. 114, no. 6, pp. 1371–1376, Feb. 2017, doi: 10.1073/pnas.1620133114.

[77] J. Dietzold, A. Gopalakrishnan, and P. Salgame, "Duality of lipid mediators in host response against Mycobacterium tuberculosis: good cop, bad cop," *F1000Prime Rep.*, vol. 7, no. March, pp. 1–8, Mar. 2015, doi: 10.12703/P7-29.

[78] E. Guirado and L. S. Schlesinger, "Modeling the Mycobacterium tuberculosis Granuloma – the Critical Battlefield in Host Immunity and Disease," *Front. Immunol.*, vol. 4, no. April, pp. 1–7, 2013, doi: 10.3389/fimmu.2013.00098.

[79] M. Silva Miranda, A. Breiman, S. Allain, F. Deknuydt, and F. Altare, "The Tuberculous Granuloma: An Unsuccessful Host Defence Mechanism Providing a Safety Shelter for the Bacteria?," *Clin. Dev. Immunol.*, vol. 2012, pp. 1–14, Jan. 2012, doi: 10.1155/2012/139127.

[80] J. E. Gomez and J. D. McKinney, "M. tuberculosis persistence, latency, and drug tolerance," *Tuberculosis*, vol. 84, no. 1–2, pp. 29–44, Jan. 2004, doi: 10.1016/j.tube.2003.08.003.

[81] N. Kapoor, S. Pawar, T. D. Sirakova, C. Deb, W. L. Warren, and P. E. Kolattukudy, "Human Granuloma In Vitro Model, for TB Dormancy and Resuscitation," *PLoS One*, vol. 8, no. 1, p. e53657, Jan. 2013, doi: 10.1371/journal.pone.0053657.

[82] C. Paige and W. R. Bishai, "Penitentiary or penthouse condo: the tuberculous granuloma from the microbe's point of view," *Cell. Microbiol.*, vol. 12, no. 3, pp. 301–309, Mar. 2010, doi: 10.1111/j.1462-5822.2009.01424.x.

[83] C. R. Shaler, C. N. Horvath, M. Jeyanathan, and Z. Xing, "Within the Enemy's Camp: contribution of the granuloma to the dissemination, persistence and transmission of Mycobacterium tuberculosis," *Front. Immunol.*, vol. 4, no. February, Jan. 2013, doi: 10.3389/fimmu.2013.00030.

[84] D. G. Russell, P.-J. Cardona, M.-J. Kim, S. Allain, and F. Altare, "Foamy macrophages and the progression of the human tuberculosis granuloma," *Nat. Immunol.*, vol. 10, no. 9, pp. 943–948, Sep. 2009, doi: 10.1038/ni.1781.

[85] R. Simeone *et al.*, "Phagosomal Rupture by Mycobacterium tuberculosis Results in Toxicity and Host Cell Death," *PLoS Pathog.*, vol. 8, no. 2, p. e1002507, Feb. 2012, doi: 10.1371/journal.ppat.1002507.

[86] M. Sani *et al.*, "Direct Visualization by Cryo-EM of the Mycobacterial Capsular Layer: A Labile Structure Containing ESX-1-Secreted Proteins," *PLoS Pathog.*, vol. 6, no. 3, p. e1000794, Mar. 2010, doi: 10.1371/journal.ppat.1000794.

[87] D. Lucarelli, M. Vasil, W. Meyer-Klaucke, and E. Pohl, "The Metal-Dependent Regulators FurA and FurB from Mycobacterium Tuberculosis," *Int. J. Mol. Sci.*, vol. 9, no. 8, pp. 1548–1560, Aug. 2008, doi: 10.3390/ijms9081548.

[88] L. J. Juttukonda and E. P. Skaar, "Manganese homeostasis and utilization in pathogenic bacteria," *Mol. Microbiol.*, vol. 97, no. 2, pp. 216–228, Jul. 2015, doi: 10.1111/mmi.13034.

[89] J. R. Forbes and P. Gros, "Iron, manganese, and cobalt transport by Nramp1 (Slc11a1) and Nramp2 (Slc11a2) expressed at the plasma membrane," *Blood*, vol. 102, no. 5, pp. 1884–1892, Sep. 2003, doi: 10.1182/blood-2003-02-0425.

[90] L. H. Sansing, T. H. Harris, F. A. Welsh, S. E. Kasner, C. A. Hunter, and K. Kariko, "Toll-like receptor 4 contributes to poor outcome after intracerebral hemorrhage," *Ann. Neurol.*, vol. 70, no. 4, pp. 646–656, Oct. 2011, doi: 10.1002/ana.22528.

[91] R. Pandey and G. M. Rodriguez, "IdeR is required for iron homeostasis and virulence in Mycobacterium tuberculosis," *Mol. Microbiol.*, vol. 91, no. 1, pp. 98–109, Jan. 2014, doi: 10.1111/mmi.12441.

[92] K. M. Papp-Wallace and M. E. Maguire, "Manganese Transport and the Role of Manganese in Virulence," *Annu. Rev. Microbiol.*, vol. 60, no. 1, pp. 187–209, Oct. 2006, doi: 10.1146/annurev.micro.60.080805.142149.

[93] S. K. Reddy, M. Kamireddi, K. Dhanireddy, L. Young, A. Davis, and P. T. Reddy, "Eukaryotic-like Adenylyl Cyclases in Mycobacterium tuberculosis H37Rv," *J. Biol. Chem.*, vol. 276, no. 37, pp. 35141–35149, Sep. 2001, doi: 10.1074/jbc.M104108200.

[94] A. Maciag *et al.*, "Global Analysis of the Mycobacterium tuberculosis Zur (FurB) Regulon," *J. Bacteriol.*, vol. 189, no. 3, pp. 730–740, Feb. 2007, doi: 10.1128/JB.01190-06.

[95] K. Kurthkoti *et al.*, "The Capacity of Mycobacterium tuberculosis To Survive Iron Starvation Might Enable It To Persist in Iron-Deprived Microenvironments of Human Granulomas," *MBio*, vol. 8, no. 4, pp. e01092-17, Sep. 2017, doi: 10.1128/mBio.01092-17.

[96] H. Lin, G. R. Andersen, and L. Yatime, "Crystal structure of human S100A8 in complex with zinc and calcium," *BMC Struct. Biol.*, vol. 16, no. 1, p. 8, Dec. 2016, doi: 10.1186/s12900-016-0058-4.

[97] O. Olakanmi, L. S. Schlesinger, A. Ahmed, and B. E. Britigan, "Intraphagosomal Mycobacterium tuberculosis Acquires Iron from Both Extracellular Transferrin and Intracellular Iron Pools," *J. Biol. Chem.*, vol. 277, no. 51, pp. 49727–49734, Dec. 2002, doi: 10.1074/jbc.M209768200.

[98] C. D. Blanchette, Y.-H. Woo, C. Thomas, N. Shen, T. A. Sulchek, and A. L. Hiddessen, "Decoupling Internalization, Acidification and Phagosomal-Endosomal/lysosomal Fusion during Phagocytosis of InlA Coated Beads in Epithelial Cells," *PLoS One*, vol. 4, no. 6, p. e6056, Jun. 2009, doi: 10.1371/journal.pone.0006056.

[99] J. R. Forbes and P. Gros, "Divalent-metal transport by NRAMP proteins at the interface of host–pathogen interactions," *Trends Microbiol.*, vol. 9, no. 8, pp. 397–403, Aug. 2001, doi: 10.1016/S0966-842X(01)02098-4.

[100] O. Olakanmi, L. S. Schlesinger, A. Ahmed, and B. E. Britigan, "The Nature of Extracellular Iron Influences Iron Acquisition by Mycobacterium tuberculosis Residing within Human Macrophages," *Infect. Immun.*, vol. 72, no. 4, pp. 2022–2028, Apr. 2004, doi: 10.1128/IAI.72.4.2022-2028.2004.

[101] N. Jabado, A. Jankowski, S. Dougaparsad, V. Picard, S. Grinstein, and P. Gros, "Natural Resistance to Intracellular Infections," *J. Exp. Med.*, vol. 192, no. 9, pp. 1237–1248, Nov. 2000, doi: 10.1084/jem.192.9.1237.

[102] F. Supek, L. Supekova, H. Nelson, and N. Nelson, "A yeast manganese transporter related to the macrophage protein involved in conferring resistance to mycobacteria.," *Proc. Natl. Acad. Sci.*, vol. 93, no. 10, pp. 5105–5110, May 1996, doi: 10.1073/pnas.93.10.5105.

[103] B. Peracino, S. Buracco, and S. Bozzaro, "The Nramp (Slc11) proteins regulate development, resistance to pathogenic bacteria and iron homeostasis in Dictyostelium discoideum," *J. Cell Sci.*, vol. 126, no. 1, pp. 301–311, Jan. 2013, doi:

10.1242/jcs.116210.

[104] D. Wagner *et al.*, "Elemental Analysis of Mycobacterium avium -, Mycobacterium tuberculosis -, and Mycobacterium smegmatis -Containing Phagosomes Indicates Pathogen-Induced Microenvironments within the Host Cell's Endosomal System," *J. Immunol.*, vol. 174, no. 3, pp. 1491–1500, Feb. 2005, doi: 10.4049/jimmunol.174.3.1491.

[105] R. Pandey, R. Russo, S. Ghanny, X. Huang, J. Helmann, and G. M. Rodriguez, "MntR(Rv2788): a transcriptional regulator that controls manganese homeostasis in M ycobacterium tuberculosis," *Mol. Microbiol.*, vol. 98, no. 6, pp. 1168–1183, Dec. 2015, doi: 10.1111/mmi.13207.

[106] E. Pohl, R. K. Holmes, and W. G. J. Hol, "Crystal Structure of the Iron-dependent Regulator (IdeR) from Mycobacteriumtuberculosis Shows Both Metal Binding Sites Fully Occupied," *J. Mol. Biol.*, vol. 285, no. 3, pp. 1145–1156, Jan. 1999, doi: 10.1006/jmbi.1998.2339.

[107] M. A. DeWitt, J. I. Kliegman, J. D. Helmann, R. G. Brennan, D. L. Farrens, and A. Glasfeld, "The Conformations of the Manganese Transport Regulator of Bacillus subtilis in its Metal-free State," *J. Mol. Biol.*, vol. 365, no. 5, pp. 1257–1265, Feb. 2007, doi: 10.1016/j.jmb.2006.10.080.

[108] M. Luo, E. A. Fadeev, and J. T. Groves, "Mycobactin-mediated iron acquisition within macrophages," *Nat. Chem. Biol.*, vol. 1, no. 3, pp. 149–153, Aug. 2005, doi: 10.1038/nchembio717.

[109] M. D. McMahon, J. S. Rush, and M. G. Thomas, "Analyses of MbtB, MbtE, and MbtF Suggest Revisions to the Mycobactin Biosynthesis Pathway in Mycobacterium tuberculosis," *J. Bacteriol.*, vol. 194, no. 11, pp. 2809–2818, Jun. 2012, doi: 10.1128/JB.00088-12.

[110] V. M. Boradia *et al.*, "Mycobacterium tuberculosis acquires iron by cell-surface sequestration and internalization of human holo-transferrin," *Nat. Commun.*, vol. 5, no. 1, p. 4730, Dec. 2014, doi: 10.1038/ncomms5730.

[111] C. A. Madigan *et al.*, "Lipidomic discovery of deoxysiderophores reveals a revised mycobactin biosynthesis pathway in Mycobacterium tuberculosis," *Proc. Natl. Acad. Sci.*, vol. 109, no. 4, pp. 1257–1262, Jan. 2012, doi: 10.1073/pnas.1109958109.

[112] P. Tyagi, A. T. Dharmaraja, A. Bhaskar, H. Chakrapani, and A. Singh, "Mycobacterium tuberculosis has diminished capacity to counteract redox stress induced by elevated levels of endogenous superoxide," *Free Radic. Biol. Med.*, vol. 84, pp. 344–354, Jul. 2015, doi: 10.1016/j.freeradbiomed.2015.03.008.

[113] G. M. Rodriguez, M. I. Voskuil, B. Gold, G. K. Schoolnik, and I. Smith, "ideR , an Essential Gene in Mycobacterium tuberculosis : Role of IdeR in Iron-Dependent Gene Expression, Iron Metabolism, and Oxidative Stress Response," *Infect. Immun.*, vol. 70, no. 7, pp. 3371–3381, Jul. 2002, doi: 10.1128/IAI.70.7.3371-3381.2002.

[114] C. Vilchèze, T. Hartman, B. Weinrick, and W. R. Jacobs, "Mycobacterium tuberculosis is extraordinarily sensitive to killing by a vitamin C-induced Fenton reaction," *Nat. Commun.*, vol. 4, no. 1, p. 1881, Oct. 2013, doi: 10.1038/ncomms2898.

[115] C. M. Litwin and S. B. Calderwood, "Role of iron in regulation of virulence genes," *Clin. Microbiol. Rev.*, vol. 6, no. 2, pp. 137–149, Apr. 1993, doi: 10.1128/CMR.6.2.137.

[116] U. E. Schaible and S. H. E. Kaufmann, "Iron and microbial infection," *Nat. Rev. Microbiol.*, vol. 2, no. 12, pp. 946–953, Dec. 2004, doi: 10.1038/nrmicro1046.

[117] F. W. Outten and E. C. Theil, "Iron-Based Redox Switches in Biology," *Antioxid. Redox Signal.*, vol. 11, no. 5, pp. 1029–1046, May 2009, doi: 10.1089/ars.2008.2296.

[118] M. S. Siegrist *et al.*, "Mycobacterial Esx-3 is required for mycobactin-mediated iron acquisition," *Proc. Natl. Acad. Sci.*, vol. 106, no. 44, pp. 18792–18797, Nov. 2009, doi: 10.1073/pnas.0900589106.

[119] A. Serafini, F. Boldrin, G. Palù, and R. Manganelli, "Characterization of a Mycobacterium tuberculosis ESX-3 Conditional Mutant: Essentiality and Rescue by Iron and Zinc," *J. Bacteriol.*, vol. 191, no. 20, pp. 6340–6344, Oct. 2009, doi: 10.1128/JB.00756-09.

[120] A. Farhana *et al.*, "Mechanistic Insights into a Novel Exporter-Importer System of Mycobacterium tuberculosis Unravel Its Role in Trafficking of Iron," *PLoS One*, vol. 3, no. 5, p. e2087, May 2008, doi: 10.1371/journal.pone.0002087.

[121] B. Gold, G. M. Rodriguez, S. A. E. Marras, M. Pentecost, and I. Smith, "The Mycobacterium tuberculosis IdeR is a dual functional regulator that controls transcription of genes involved in iron acquisition, iron storage and survival in macrophages," *Mol. Microbiol.*, vol. 42, no. 3, pp. 851–865, Jul. 2008, doi: 10.1046/j.1365-2958.2001.02684.x.

[122] G. M. Rodriguez and I. Smith, "Mechanisms of iron regulation in mycobacteria: role in physiology and virulence," *Mol. Microbiol.*, vol. 47, no. 6, pp. 1485–1494, Mar. 2003, doi: 10.1046/j.1365-2958.2003.03384.x.

[123] D. A. Oldridge *et al.*, "Genetic predisposition to neuroblastoma mediated by a LMO1 super-enhancer polymorphism," *Nature*, vol. 528, no. 7582, pp. 418–421, Dec. 2015, doi: 10.1038/nature15540.

[124] G. Fu *et al.*, "Correction for Fu et al., Female-specific flightless phenotype for mosquito control," *Proc. Natl. Acad. Sci.*, vol. 107, no. 43, pp. 18741–18741, Oct. 2010, doi: 10.1073/pnas.1014662107.

[125] R. Colangeli *et al.*, "The multifunctional histone-like protein Lsr2 protects mycobacteria against reactive oxygen intermediates," *Proc. Natl. Acad. Sci.*, vol. 106, no. 11, pp. 4414–4418, Mar. 2009, doi: 10.1073/pnas.0810126106.

[126] R. Colangeli *et al.*, "Transcriptional Regulation of Multi-Drug Tolerance and Antibiotic-Induced Responses by the Histone-Like Protein Lsr2 in M. tuberculosis," *PLoS Pathog.*, vol. 3, no. 6, p. e87, Jun. 2007, doi: 10.1371/journal.ppat.0030087.

[127] G. Khare, P. Nangpal, and A. K. Tyagi, "Differential Roles of Iron Storage Proteins in Maintaining the Iron Homeostasis in Mycobacterium tuberculosis," *PLoS One*, vol. 12, no. 1, p. e0169545, Jan. 2017, doi: 10.1371/journal.pone.0169545.

[128] C. M. Jones *et al.*, "Self-poisoning of Mycobacterium tuberculosis by interrupting siderophore recycling," *Proc. Natl. Acad. Sci.*, vol. 111, no. 5, pp. 1945–1950, Feb. 2014, doi: 10.1073/pnas.1311402111.

[129] S. D. Pandey *et al.*, "Iron-Regulated Protein HupB of Mycobacterium tuberculosis Positively Regulates Siderophore Biosynthesis and Is Essential for Growth in Macrophages," *J. Bacteriol.*, vol. 196, no. 10, pp. 1853–1865, May 2014, doi: 10.1128/JB.01483-13.

[130] M. Sritharan, "Iron Homeostasis in Mycobacterium tuberculosis: Mechanistic Insights into Siderophore-Mediated Iron Uptake," *J. Bacteriol.*, vol. 198, no. 18, pp. 2399–2409, Sep. 2016, doi: 10.1128/JB.00359-16.

[131] J. Gonzalo-Asensio *et al.*, "PhoP: A Missing Piece in the Intricate Puzzle of

Mycobacterium tuberculosis Virulence," *PLoS One*, vol. 3, no. 10, p. e3496, Oct. 2008, doi: 10.1371/journal.pone.0003496.

[132] S. Gupta, A. Sinha, and D. Sarkar, "Transcriptional autoregulation by Mycobacterium tuberculosis PhoP involves recognition of novel direct repeat sequences in the regulatory region of the promoter," *FEBS Lett.*, vol. 580, no. 22, pp. 5328–5338, Oct. 2006, doi: 10.1016/j.febslet.2006.09.004.

[133] J. Gonzalo-Asensio *et al.*, "The Mycobacterium tuberculosis phoPR Operon Is Positively Autoregulated in the Virulent Strain H37Rv," *J. Bacteriol.*, vol. 190, no. 21, pp. 7068–7078, Nov. 2008, doi: 10.1128/JB.00712-08.

[134] Y. C. Manabe, B. J. Saviola, L. Sun, J. R. Murphy, and W. R. Bishai, "Attenuation of virulence in Mycobacterium tuberculosis expressing a constitutively active iron repressor," *Proc. Natl. Acad. Sci.*, vol. 96, no. 22, pp. 12844–12848, Oct. 1999, doi: 10.1073/pnas.96.22.12844.

[135] S. Banerjee, A. K. Nandyala, P. Raviprasad, N. Ahmed, and S. E. Hasnain, "Iron-Dependent RNA-Binding Activity of Mycobacterium tuberculosis Aconitase," *J. Bacteriol.*, vol. 189, no. 11, pp. 4046–4052, Jun. 2007, doi: 10.1128/JB.00026-07.

[136] K. B. Pechter, F. M. Meyer, A. W. Serio, J. Stülke, and A. L. Sonenshein, "Two Roles for Aconitase in the Regulation of Tricarboxylic Acid Branch Gene Expression in Bacillus subtilis," *J. Bacteriol.*, vol. 195, no. 7, pp. 1525–1537, Apr. 2013, doi: 10.1128/JB.01690-12.

[137] K. H. Rohde, R. B. Abramovitch, and D. G. Russell, "Mycobacterium tuberculosis Invasion of Macrophages: Linking Bacterial Gene Expression to Environmental Cues," *Cell Host Microbe*, vol. 2, no. 5, pp. 352–364, Nov. 2007, doi: 10.1016/j.chom.2007.09.006.

[138] Y. Hu *et al.*, "σ E -dependent activation of RbpA controls transcription of the furA-katG operon in response to oxidative stress in mycobacteria," *Mol. Microbiol.*, vol. 102, no. 1, pp. 107–120, Oct. 2016, doi: 10.1111/mmi.13449.

[139] F. Zheng, Q. Long, and J. Xie, "The Function and Regulatory Network of WhiB and WhiB-Like Protein from Comparative Genomics and Systems Biology Perspectives," *Cell Biochem. Biophys.*, vol. 63, no. 2, pp. 103–108, Jun. 2012, doi: 10.1007/s12013-012-9348-z.

[140] A. Z. Reeves *et al.*, "Aminoglycoside Cross-Resistance in Mycobacterium tuberculosis Due to Mutations in the 5′ Untranslated Region of whiB7," *Antimicrob. Agents Chemother.*, vol. 57, no. 4, pp. 1857–1865, Apr. 2013, doi: 10.1128/AAC.02191-12.

[141] L. P. Samuel *et al.*, "Expression, production and release of the Eis protein by Mycobacterium tuberculosis during infection of macrophages and its effect on cytokine secretion," *Microbiology*, vol. 153, no. 2, pp. 529–540, Feb. 2007, doi: 10.1099/mic.0.2006/002642-0.

[142] M. Farina, D. S. Avila, J. B. T. da Rocha, and M. Aschner, "Metals, oxidative stress and neurodegeneration: A focus on iron, manganese and mercury," *Neurochem. Int.*, vol. 62, no. 5, pp. 575–594, Apr. 2013, doi: 10.1016/j.neuint.2012.12.006.

[143] O. L. Champion *et al.*, "Yersinia pseudotuberculosis mntH functions in intracellular manganese accumulation, which is essential for virulence and survival in cells expressing functional Nramp1," *Microbiology*, vol. 157, no. 4, pp. 1115–1122, Apr. 2011, doi: 10.1099/mic.0.045807-0.

[144] A. Kumar, A. Farhana, L. Guidry, V. Saini, M. Hondalus, and A. J. C. Steyn, "Redox homeostasis in mycobacteria: the key to tuberculosis control?," *Expert Rev. Mol.*

Med., vol. 13, no. December, p. e39, Dec. 2011, doi: 10.1017/S1462399411002079.

[145] Y. Akhter, S. Yellaboina, A. Farhana, A. Ranjan, N. Ahmed, and S. E. Hasnain, "Genome scale portrait of cAMP-receptor protein (CRP) regulons in mycobacteria points to their role in pathogenesis," *Gene*, vol. 407, no. 1–2, pp. 148–158, Jan. 2008, doi: 10.1016/j.gene.2007.10.017.

[146] N. Matange, "Revisiting bacterial cyclic nucleotide phosphodiesterases: cyclic AMP hydrolysis and beyond," *FEMS Microbiol. Lett.*, vol. 362, no. 22, p. fnv183, Nov. 2015, doi: 10.1093/femsle/fnv183.

[147] B. K. M. Dass, R. Sharma, A. R. Shenoy, R. Mattoo, and S. S. Visweswariah, "Cyclic AMP in Mycobacteria: Characterization and Functional Role of the Rv1647 Ortholog in Mycobacterium smegmatis," *J. Bacteriol.*, vol. 190, no. 11, pp. 3824–3834, Jun. 2008, doi: 10.1128/JB.00138-08.

[148] G. Bai, G. S. Knapp, and K. A. McDonough, "Cyclic AMP signalling in mycobacteria: redirecting the conversation with a common currency," *Cell. Microbiol.*, vol. 13, no. 3, pp. 349–358, Mar. 2011, doi: 10.1111/j.1462-5822.2010.01562.x.

[149] D. Dittrich, C. Keller, S. Ehlers, J. E. Schultz, and P. Sander, "Characterization of a Mycobacterium tuberculosis mutant deficient in pH-sensing adenylate cyclase Rv1264," *Int. J. Med. Microbiol.*, vol. 296, no. 8, pp. 563–566, Dec. 2006, doi: 10.1016/j.ijmm.2006.07.001.

[150] N. Agarwal, G. Lamichhane, R. Gupta, S. Nolan, and W. R. Bishai, "Cyclic AMP intoxication of macrophages by a Mycobacterium tuberculosis adenylate cyclase," *Nature*, vol. 460, no. 7251, pp. 98–102, Jul. 2009, doi: 10.1038/nature08123.

[151] J. Daniel, L. Abraham, A. Martin, X. Pablo, and S. Reyes, "Rv2477c is an antibiotic-sensitive manganese-dependent ABC-F ATPase in Mycobacterium tuberculosis," *Biochem. Biophys. Res. Commun.*, vol. 495, no. 1, pp. 35–40, Jan. 2018, doi: 10.1016/j.bbrc.2017.10.168.

[152] P. Rishi, N. Jindal, S. Bharrhan, and R. P. Tiwari, "Salmonella–Macrophage Interactions upon Manganese Supplementation," *Biol. Trace Elem. Res.*, vol. 133, no. 1, pp. 110–119, Jan. 2010, doi: 10.1007/s12011-009-8406-x.

[153] V. E. Diaz-Ochoa *et al.*, "Salmonella Mitigates Oxidative Stress and Thrives in the Inflamed Gut by Evading Calprotectin-Mediated Manganese Sequestration," *Cell Host Microbe*, vol. 19, no. 6, pp. 814–825, Jun. 2016, doi: 10.1016/j.chom.2016.05.005.

[154] D. Agranoff, I. M. Monahan, J. A. Mangan, P. D. Butcher, and S. Krishna, "Mycobacterium tuberculosis Expresses a Novel Ph-Dependent Divalent Cation Transporter Belonging to the Nramp Family," *J. Exp. Med.*, vol. 190, no. 5, pp. 717–724, Sep. 1999, doi: 10.1084/jem.190.5.717.

[155] X. Pang *et al.*, "MprAB Regulates the espA Operon in Mycobacterium tuberculosis and Modulates ESX-1 Function and Host Cytokine Response," *J. Bacteriol.*, vol. 195, no. 1, pp. 66–75, Jan. 2013, doi: 10.1128/JB.01067-12.

[156] Z. Chen *et al.*, "Mycobacterial WhiB6 Differentially Regulates ESX-1 and the Dos Regulon to Modulate Granuloma Formation and Virulence in Zebrafish," *Cell Rep.*, vol. 16, no. 9, pp. 2512–2524, Aug. 2016, doi: 10.1016/j.celrep.2016.07.080.

[157] S. B. Korch, H. Contreras, and J. E. Clark-Curtiss, "Three Mycobacterium tuberculosis Rel Toxin-Antitoxin Modules Inhibit Mycobacterial Growth and Are Expressed in Infected Human Macrophages," *J. Bacteriol.*, vol. 191, no. 5, pp. 1618–1630, Mar. 2009, doi: 10.1128/JB.01318-08.

[158] M. Yang, C.-H. Gao, J. Hu, C. Dong, and Z.-G. He, "Characterization of the interaction between a SirR family transcriptional factor of Mycobacterium tuberculosis , encoded by Rv2788, and a pair of toxin-antitoxin proteins RelJ/K, encoded by Rv3357 and Rv3358," *FEBS J.*, vol. 281, no. 12, pp. 2726–2737, Jun. 2014, doi: 10.1111/febs.12815.

[159] S.-M. Kang *et al.*, "Functional details of the Mycobacterium tuberculosis VapBC26 toxin-antitoxin system based on a structural study: insights into unique binding and antibiotic peptides," *Nucleic Acids Res.*, vol. 45, no. 14, pp. 8564–8580, Aug. 2017, doi: 10.1093/nar/gkx489.

[160] I.-G. Lee, S. J. Lee, S. Chae, K.-Y. Lee, J. Kim, and B. Lee, "Structural and functional studies of the Mycobacterium tuberculosis VapBC30 toxin-antitoxin system: implications for the design of novel antimicrobial peptides," *Nucleic Acids Res.*, vol. 43, no. 15, pp. 7624–7637, Sep. 2015, doi: 10.1093/nar/gkv689.

[161] A. Serafini, D. Pisu, G. Palù, G. M. Rodriguez, and R. Manganelli, "The ESX-3 Secretion System Is Necessary for Iron and Zinc Homeostasis in Mycobacterium tuberculosis," *PLoS One*, vol. 8, no. 10, p. e78351, Oct. 2013, doi: 10.1371/journal.pone.0078351.

[162] D. J. Bretl, T. M. Bigley, S. S. Terhune, and T. C. Zahrt, "The MprB Extracytoplasmic Domain Negatively Regulates Activation of the Mycobacterium tuberculosis MprAB Two-Component System," *J. Bacteriol.*, vol. 196, no. 2, pp. 391–406, Jan. 2014, doi: 10.1128/JB.01064-13.

[163] D. Ilghari *et al.*, "Solution Structure of the Mycobacterium tuberculosis EsxG·EsxH Complex," *J. Biol. Chem.*, vol. 286, no. 34, pp. 29993–30002, Aug. 2011, doi: 10.1074/jbc.M111.248732.

[164] S. E. Gabriel and J. D. Helmann, "Contributions of Zur-Controlled Ribosomal Proteins to Growth under Zinc Starvation Conditions," *J. Bacteriol.*, vol. 191, no. 19, pp. 6116–6122, Oct. 2009, doi: 10.1128/JB.00802-09.

[165] S. Seto, K. Tsujimura, and Y. Koide, "Rab GTPases Regulating Phagosome Maturation Are Differentially Recruited to Mycobacterial Phagosomes," *Traffic*, vol. 12, no. 4, pp. 407–420, Apr. 2011, doi: 10.1111/j.1600-0854.2011.01165.x.

[166] A. Mehra *et al.*, "Mycobacterium tuberculosis Type VII Secreted Effector EsxH Targets Host ESCRT to Impair Trafficking," *PLoS Pathog.*, vol. 9, no. 10, p. e1003734, Oct. 2013, doi: 10.1371/journal.ppat.1003734.

[167] L. E. Via, D. Deretic, R. J. Ulmer, N. S. Hibler, L. A. Huber, and V. Deretic, "Arrest of Mycobacterial Phagosome Maturation Is Caused by a Block in Vesicle Fusion between Stages Controlled by rab5 and rab7," *J. Biol. Chem.*, vol. 272, no. 20, pp. 13326–13331, May 1997, doi: 10.1074/jbc.272.20.13326.

[168] D. Wong, H. Bach, J. Sun, Z. Hmama, and Y. Av-Gay, "Mycobacterium tuberculosis protein tyrosine phosphatase (PtpA) excludes host vacuolar-H+-ATPase to inhibit phagosome acidification.," *PNAS*, vol. 108, no. 48, pp. 19371–19376, Nov. 2011, doi: 10.1073/pnas.1109201108.

[169] S. a Kalamidas *et al.*, "cAMP synthesis and degradation by phagosomes regulate actin assembly and fusion events: consequences for mycobacteria," *J. Cell Sci.*, vol. 119, no. 17, pp. 3686–3694, Sep. 2006, doi: 10.1242/jcs.03091.

[170] A. Gupta, A. Kaul, A. G. Tsolaki, U. Kishore, and S. Bhakta, "Mycobacterium tuberculosis: Immune evasion, latency and reactivation," *Immunobiology*, vol. 217, no. 3, pp. 363–374, Mar. 2012, doi: 10.1016/j.imbio.2011.07.008.

[171] E. P. Thi, U. Lambertz, and N. E. Reiner, "Sleeping with the Enemy: How Intracellular

Pathogens Cope with a Macrophage Lifestyle," *PLoS Pathog.*, vol. 8, no. 3, p. e1002551, Mar. 2012, doi: 10.1371/journal.ppat.1002551.

[172] L. Nguyen and J. Pieters, "Mycobacterial Subversion of Chemotherapeutic Reagents and Host Defense Tactics: Challenges in Tuberculosis Drug Development," *Annu. Rev. Pharmacol. Toxicol.*, vol. 49, no. 1, pp. 427–453, Feb. 2009, doi: 10.1146/annurev-pharmtox-061008-103123.

[173] M. S. Alam, S. K. Garg, and P. Agrawal, "Studies on structural and functional divergence among seven WhiB proteins of Mycobacterium tuberculosis H37Rv," *FEBS J.*, vol. 276, no. 1, pp. 76–93, Jan. 2009, doi: 10.1111/j.1742-4658.2008.06755.x.

[174] C. Larsson, B. Luna, N. C. Ammerman, M. Maiga, N. Agarwal, and W. R. Bishai, "Gene Expression of Mycobacterium tuberculosis Putative Transcription Factors whiB1-7 in Redox Environments," *PLoS One*, vol. 7, no. 7, p. e37516, Jul. 2012, doi: 10.1371/journal.pone.0037516.

[175] L. J. Smith *et al.*, "Mycobacterium tuberculosis WhiB1 is an essential DNA-binding protein with a nitric oxide-sensitive iron–sulfur cluster," *Biochem. J.*, vol. 432, no. 3, pp. 417–427, Dec. 2010, doi: 10.1042/BJ20101440.

[176] S. Ranganathan *et al.*, "Characterization of a cAMP responsive transcription factor, Cmr (Rv1675c), in TB complex mycobacteria reveals overlap with the DosR (DevR) dormancy regulon," *Nucleic Acids Res.*, vol. 44, no. 1, pp. 134–151, Jan. 2016, doi: 10.1093/nar/gkv889.

[177] M. Chawla *et al.*, "Mycobacterium tuberculosis WhiB4 regulates oxidative stress response to modulate survival and dissemination in vivo," *Mol. Microbiol.*, vol. 85, no. 6, pp. 1148–1165, Sep. 2012, doi: 10.1111/j.1365-2958.2012.08165.x.

[178] S. Casonato *et al.*, "WhiB5, a Transcriptional Regulator That Contributes to Mycobacterium tuberculosis Virulence and Reactivation," *Infect. Immun.*, vol. 80, no. 9, pp. 3132–3144, Sep. 2012, doi: 10.1128/IAI.06328-11.

[179] A. Singh *et al.*, "Mycobacterium tuberculosis WhiB3 responds to O 2 and nitric oxide via its [4Fe-4S] cluster and is essential for nutrient starvation survival," *Proc. Natl. Acad. Sci.*, vol. 104, no. 28, pp. 11562–11567, Jul. 2007, doi: 10.1073/pnas.0700490104.

[180] M. R. Stapleton, L. J. Smith, D. M. Hunt, R. S. Buxton, and J. Green, "Mycobacterium tuberculosis WhiB1 represses transcription of the essential chaperonin GroEL2," *Tuberculosis*, vol. 92, no. 4, pp. 328–332, Jul. 2012, doi: 10.1016/j.tube.2012.03.001.

[181] R. B. Abramovitch, K. H. Rohde, F.-F. Hsu, and D. G. Russell, "aprABC: a Mycobacterium tuberculosis complex-specific locus that modulates pH-driven adaptation to the macrophage phagosome," *Mol. Microbiol.*, vol. 80, no. 3, pp. 678–694, May 2011, doi: 10.1111/j.1365-2958.2011.07601.x.

[182] A. Singh *et al.*, "Mycobacterium tuberculosis WhiB3 Maintains Redox Homeostasis by Regulating Virulence Lipid Anabolism to Modulate Macrophage Response," *PLoS Pathog.*, vol. 5, no. 8, p. e1000545, Aug. 2009, doi: 10.1371/journal.ppat.1000545.

[183] R. A. Rienksma *et al.*, "Comprehensive insights into transcriptional adaptation of intracellular mycobacteria by microbe-enriched dual RNA sequencing," *BMC Genomics*, vol. 16, no. 1, p. 34, Dec. 2015, doi: 10.1186/s12864-014-1197-2.

[184] M. Zimmermann *et al.*, "Integration of Metabolomics and Transcriptomics Reveals a Complex Diet of Mycobacterium tuberculosis during Early Macrophage Infection," *mSystems*, vol. 2, no. 4, pp. 1–18, Aug. 2017, doi: 10.1128/mSystems.00057-17.

[185] V. Saini, A. Farhana, and A. J. C. Steyn, "Mycobacterium tuberculosis WhiB3: A Novel Iron–Sulfur Cluster Protein That Regulates Redox Homeostasis and Virulence," *Antioxid. Redox Signal.*, vol. 16, no. 7, pp. 687–697, Apr. 2012, doi: 10.1089/ars.2011.4341.

[186] M. A. Gazdik and K. A. McDonough, "Identification of Cyclic AMP-Regulated Genes in Mycobacterium tuberculosis Complex Bacteria under Low-Oxygen Conditions," *J. Bacteriol.*, vol. 187, no. 8, pp. 2681–2692, Apr. 2005, doi: 10.1128/JB.187.8.2681-2692.2005.

[187] S. Joseph, A. Yuen, V. Singh, and Z. Hmama, "Mycobacterium tuberculosis Cpn60.2 (GroEL2) blocks macrophage apoptosis via interaction with mitochondrial mortalin," *Biol. Open*, vol. 2, no. 6, pp. 481–488, Jan. 2017, doi: 10.1242/bio.023119.

[188] J. L. Naffin-Olivos *et al.*, "Mycobacterium tuberculosis Hip1 Modulates Macrophage Responses through Proteolysis of GroEL2," *PLoS Pathog.*, vol. 10, no. 5, p. e1004132, May 2014, doi: 10.1371/journal.ppat.1004132.

[189] M. A. Forrellad *et al.*, "Virulence factors of the Mycobacterium tuberculosis complex," *Virulence*, vol. 4, no. 1, pp. 3–66, Jan. 2013, doi: 10.4161/viru.22329.

[190] S. Jamwal, M. K. Midha, H. N. Verma, A. Basu, K. V. S. Rao, and V. Manivel, "Characterizing virulence-specific perturbations in the mitochondrial function of macrophages infected with mycobacterium tuberculosis," *Sci. Rep.*, vol. 3, no. 1, p. 1328, Dec. 2013, doi: 10.1038/srep01328.

[191] G. R. Stewart *et al.*, "Dissection of the heat-shock response in Mycobacterium tuberculosis using mutants and microarrays a aA list of the 100 ORFs most highly induced by heat shock is provided as supplementary data with the online version of this paper (http://mic.sgmjournals.o," *Microbiology*, vol. 148, no. 10, pp. 3129–3138, Oct. 2002, doi: 10.1099/00221287-148-10-3129.

[192] N. Agarwal, T. R. Raghunand, and W. R. Bishai, "Regulation of the expression of whiB1 in Mycobacterium tuberculosis: role of cAMP receptor protein," *Microbiology*, vol. 152, no. 9, pp. 2749–2756, Sep. 2006, doi: 10.1099/mic.0.28924-0.

[193] M. A. Gazdik, G. Bai, Y. Wu, and K. A. McDonough, "Rv1675c (cmr) regulates intramacrophage and cyclic AMP-induced gene expression in Mycobacterium tuberculosis -complex mycobacteria," *Mol. Microbiol.*, vol. 71, no. 2, pp. 434–448, Jan. 2009, doi: 10.1111/j.1365-2958.2008.06541.x.

[194] E. N. G. Houben, K. V. Korotkov, and W. Bitter, "Take five — Type VII secretion systems of Mycobacteria," *Biochim. Biophys. Acta - Mol. Cell Res.*, vol. 1843, no. 8, pp. 1707–1716, Aug. 2014, doi: 10.1016/j.bbamcr.2013.11.003.

[195] M. Newton-Foot, "The Mycobacterium tuberculosis ESX-3 secretion system interactome," 2010. [Online]. Available: http://scholar.sun.ac.za/handle/10019.1/4841

[196] R. Simeone, D. Bottai, and R. Brosch, "ESX/type VII secretion systems and their role in host–pathogen interaction," *Curr. Opin. Microbiol.*, vol. 12, no. 1, pp. 4–10, Feb. 2009, doi: 10.1016/j.mib.2008.11.003.

[197] J. M. Tufariello *et al.*, "Separable roles for Mycobacterium tuberculosis ESX-3 effectors in iron acquisition and virulence," *Proc. Natl. Acad. Sci.*, vol. 113, no. 3, p. 201523321, Jan. 2016, doi: 10.1073/pnas.1523321113.

[198] B. M. Tiwari, N. Kannan, L. Vemu, and T. R. Raghunand, "The Mycobacterium tuberculosis PE Proteins Rv0285 and Rv1386 Modulate Innate Immunity and Mediate Bacillary Survival in Macrophages," *PLoS One*, vol. 7, no. 12, p. e51686, Dec.

2012, doi: 10.1371/journal.pone.0051686.

[199] W. Li, Q. Zhao, W. Deng, T. Chen, M. Liu, and J. Xie, "Mycobacterium tuberculosis Rv3402c Enhances Mycobacterial Survival within Macrophages and Modulates the Host Pro-Inflammatory Cytokines Production via NF-Kappa B/ERK/p38 Signaling," *PLoS One*, vol. 9, no. 4, p. e94418, Apr. 2014, doi: 10.1371/journal.pone.0094418.

[200] S. Daim *et al.*, "Expression of the Mycobacterium tuberculosis PPE37 protein in Mycobacterium smegmatis induces low tumour necrosis factor alpha and interleukin 6 production in murine macrophages," *J. Med. Microbiol.*, vol. 60, no. 5, pp. 582–591, May 2011, doi: 10.1099/jmm.0.026047-0.

[201] V. C. Yeruva, A. Kulkarni, R. Khandelwal, Y. Sharma, and T. R. Raghunand, "The PE_PGRS Proteins of Mycobacterium tuberculosis Are Ca 2+ Binding Mediators of Host–Pathogen Interaction," *Biochemistry*, vol. 55, no. 33, pp. 4675–4687, Aug. 2016, doi: 10.1021/acs.biochem.6b00289.

[202] L. Meng *et al.*, "PPE38 Protein of Mycobacterium tuberculosis Inhibits Macrophage MHC Class I Expression and Dampens CD8+ T Cell Responses," *Front. Cell. Infect. Microbiol.*, vol. 7, no. March, pp. 1–11, Mar. 2017, doi: 10.3389/fcimb.2017.00068.

[203] D. Dong *et al.*, "PPE38 Modulates the Innate Immune Response and Is Required for Mycobacterium marinum Virulence," *Infect. Immun.*, vol. 80, no. 1, pp. 43–54, Jan. 2012, doi: 10.1128/IAI.05249-11.

[204] J. C. Cyktor, B. Carruthers, R. A. Kominsky, G. L. Beamer, P. Stromberg, and J. Turner, "IL-10 Inhibits Mature Fibrotic Granuloma Formation during Mycobacterium tuberculosis Infection," *J. Immunol.*, vol. 190, no. 6, pp. 2778–2790, Mar. 2013, doi: 10.4049/jimmunol.1202722.

[205] S. Mahajan *et al.*, "Mycobacterium tuberculosis Modulates Macrophage Lipid-Sensing Nuclear Receptors PPARγ and TR4 for Survival," *J. Immunol.*, vol. 188, no. 11, pp. 5593–5603, Jun. 2012, doi: 10.4049/jimmunol.1103038.

[206] X. Han, S. Kitamoto, H. Wang, and W. A. Boisvert, "Interleukin-10 overexpression in macrophages suppresses atherosclerosis in hyperlipidemic mice," *FASEB J.*, vol. 24, no. 8, pp. 2869–2880, Aug. 2010, doi: 10.1096/fj.09-148155.

[207] R. Prados-Rosales, B. C. Weinrick, D. G. Piqué, W. R. Jacobs, A. Casadevall, and G. M. Rodriguez, "Role for Mycobacterium tuberculosis Membrane Vesicles in Iron Acquisition," *J. Bacteriol.*, vol. 196, no. 6, pp. 1250–1256, Mar. 2014, doi: 10.1128/JB.01090-13.

[208] J. Smith *et al.*, "Evidence for Pore Formation in Host Cell Membranes by ESX-1-Secreted ESAT-6 and Its Role in Mycobacterium marinum Escape from the Vacuole," *Infect. Immun.*, vol. 76, no. 12, pp. 5478–5487, Dec. 2008, doi: 10.1128/IAI.00614-08.

[209] Y.-J. Lim *et al.*, "Mycobacterium kansasii-induced death of murine macrophages involves endoplasmic reticulum stress responses mediated by reactive oxygen species generation or calpain activation," *Apoptosis*, vol. 18, no. 2, pp. 150–159, Feb. 2013, doi: 10.1007/s10495-012-0792-4.

[210] F. Mba Medie, M. M. Champion, E. A. Williams, and P. A. D. Champion, "Homeostasis of N-α-Terminal Acetylation of EsxA Correlates with Virulence in Mycobacterium marinum," *Infect. Immun.*, vol. 82, no. 11, pp. 4572–4586, Nov. 2014, doi: 10.1128/IAI.02153-14.

[211] J. Augenstreich *et al.*, "ESX-1 and phthiocerol dimycocerosates of Mycobacterium tuberculosis act in concert to cause phagosomal rupture and host cell apoptosis," *Cell.*

Microbiol., vol. 19, no. 7, p. e12726, Jul. 2017, doi: 10.1111/cmi.12726.

[212] R. J. Francis, R. E. Butler, and G. R. Stewart, "Mycobacterium tuberculosis ESAT-6 is a leukocidin causing Ca2+ influx, necrosis and neutrophil extracellular trap formation," *Cell Death Dis.*, vol. 5, no. 10, pp. e1474–e1474, Oct. 2014, doi: 10.1038/cddis.2014.394.

[213] H. S. Clemmensen *et al.*, "An attenuated Mycobacterium tuberculosis clinical strain with a defect in ESX-1 secretion induces minimal host immune responses and pathology," *Sci. Rep.*, vol. 7, no. 1, p. 46666, May 2017, doi: 10.1038/srep46666.

[214] W. Deng, X. Xiang, and J. Xie, "Comparative Genomic and Proteomic Anatomy of Mycobacterium Ubiquitous Esx Family Proteins: Implications in Pathogenicity and Virulence," *Curr. Microbiol.*, vol. 68, no. 4, pp. 558–567, Apr. 2014, doi: 10.1007/s00284-013-0507-2.

[215] P. A. DiGiuseppe Champion, M. M. Champion, P. Manzanillo, and J. S. Cox, "ESX-1 secreted virulence factors are recognized by multiple cytosolic AAA ATPases in pathogenic mycobacteria," *Mol. Microbiol.*, vol. 73, no. 5, pp. 950–962, Sep. 2009, doi: 10.1111/j.1365-2958.2009.06821.x.

[216] J. M. Chen *et al.*, "EspD Is Critical for the Virulence-Mediating ESX-1 Secretion System in Mycobacterium tuberculosis," *J. Bacteriol.*, vol. 194, no. 4, pp. 884–893, Feb. 2012, doi: 10.1128/JB.06417-11.

[217] B. Ize and T. Palmer, "Mycobacteria's export strategy.," *Science*, vol. 313, no. 5793, pp. 1583–4, Sep. 2006, doi: 10.1126/science.1132537.

[218] S. M. Fortune *et al.*, "Mutually dependent secretion of proteins required for mycobacterial virulence," *Proc. Natl. Acad. Sci.*, vol. 102, no. 30, pp. 10676–10681, Jul. 2005, doi: 10.1073/pnas.0504922102.

[219] A. Garces *et al.*, "EspA Acts as a Critical Mediator of ESX1-Dependent Virulence in Mycobacterium tuberculosis by Affecting Bacterial Cell Wall Integrity," *PLoS Pathog.*, vol. 6, no. 6, p. e1000957, Jun. 2010, doi: 10.1371/journal.ppat.1000957.

[220] L. S. Ates and R. Brosch, "Discovery of the type VII ESX-1 secretion needle?," *Mol. Microbiol.*, vol. 103, no. 1, pp. 7–12, Jan. 2017, doi: 10.1111/mmi.13579.

[221] Y. Lou, J. Rybniker, C. Sala, and S. T. Cole, "EspC forms a filamentous structure in the cell envelope of M ycobacterium tuberculosis and impacts ESX-1 secretion," *Mol. Microbiol.*, vol. 103, no. 1, pp. 26–38, Jan. 2017, doi: 10.1111/mmi.13575.

[222] Y. M. Ohol, D. H. Goetz, K. Chan, M. U. Shiloh, C. S. Craik, and J. S. Cox, "Mycobacterium tuberculosis MycP1 Protease Plays a Dual Role in Regulation of ESX-1 Secretion and Virulence," *Cell Host Microbe*, vol. 7, no. 3, pp. 210–220, Mar. 2010, doi: 10.1016/j.chom.2010.02.006.

[223] V. J. C. van Winden *et al.*, "Mycosins Are Required for the Stabilization of the ESX-1 and ESX-5 Type VII Secretion Membrane Complexes," *MBio*, vol. 7, no. 5, pp. 1–11, Nov. 2016, doi: 10.1128/mBio.01471-16.

[224] A. Sinha, S. Gupta, S. Bhutani, A. Pathak, and D. Sarkar, "PhoP-PhoP Interaction at Adjacent PhoP Binding Sites Is Influenced by Protein Phosphorylation," *J. Bacteriol.*, vol. 190, no. 4, pp. 1317–1328, Feb. 2008, doi: 10.1128/JB.01074-07.

[225] E. Pérez, S. Samper, Y. Bordas, C. Guilhot, B. Gicquel, and C. Martín, "An essential role for phoP in Mycobacterium tuberculosis virulence," *Mol. Microbiol.*, vol. 41, no. 1, pp. 179–187, Dec. 2001, doi: 10.1046/j.1365-2958.2001.02500.x.

[226] R. Bansal, V. Anil Kumar, R. R. Sevalkar, P. R. Singh, and D. Sarkar, "Mycobacterium

tuberculosis virulence-regulator PhoP interacts with alternative sigma factor SigE during acid-stress response," *Mol. Microbiol.*, vol. 104, no. 3, pp. 400–411, May 2017, doi: 10.1111/mmi.13635.

[227] R. Singh, M. Singh, G. Arora, S. Kumar, P. Tiwari, and S. Kidwai, "Polyphosphate Deficiency in Mycobacterium tuberculosis Is Associated with Enhanced Drug Susceptibility and Impaired Growth in Guinea Pigs," *J. Bacteriol.*, vol. 195, no. 12, pp. 2839–2851, Jun. 2013, doi: 10.1128/JB.00038-13.

[228] K. Sureka *et al.*, "Polyphosphate kinase is involved in stress-induced mprAB-sigE-rel signalling in mycobacteria," *Mol. Microbiol.*, vol. 65, no. 2, pp. 261–276, Jul. 2007, doi: 10.1111/j.1365-2958.2007.05814.x.

[229] D. J. Bretl, C. Demetriadou, and T. C. Zahrt, "Adaptation to Environmental Stimuli within the Host: Two-Component Signal Transduction Systems of Mycobacterium tuberculosis," *Microbiol. Mol. Biol. Rev.*, vol. 75, no. 4, pp. 566–582, Dec. 2011, doi: 10.1128/MMBR.05004-11.

[230] V. Anil Kumar *et al.*, "EspR-dependent ESAT-6 Protein Secretion of Mycobacterium tuberculosis Requires the Presence of Virulence Regulator PhoP," *J. Biol. Chem.*, vol. 291, no. 36, pp. 19018–19030, Sep. 2016, doi: 10.1074/jbc.M116.746289.

[231] M. I. Gröschel, F. Sayes, R. Simeone, L. Majlessi, and R. Brosch, "ESX secretion systems: mycobacterial evolution to counter host immunity," *Nat. Rev. Microbiol.*, vol. 14, no. 11, pp. 677–691, Nov. 2016, doi: 10.1038/nrmicro.2016.131.

[232] W. Bitter *et al.*, "Systematic Genetic Nomenclature for Type VII Secretion Systems," *PLoS Pathog.*, vol. 5, no. 10, p. e1000507, Oct. 2009, doi: 10.1371/journal.ppat.1000507.

[233] C. Kahramanoglou *et al.*, "Genomic mapping of cAMP receptor protein (CRP Mt) in Mycobacterium tuberculosis : relation to transcriptional start sites and the role of CRP Mt as a transcription factor," *Nucleic Acids Res.*, vol. 42, no. 13, pp. 8320–8329, Jul. 2014, doi: 10.1093/nar/gku548.

[234] M. I. de Jonge *et al.*, "ESAT-6 from Mycobacterium tuberculosis Dissociates from Its Putative Chaperone CFP-10 under Acidic Conditions and Exhibits Membrane-Lysing Activity," *J. Bacteriol.*, vol. 189, no. 16, pp. 6028–6034, Aug. 2007, doi: 10.1128/JB.00469-07.

[235] B. Blasco *et al.*, "Virulence Regulator EspR of Mycobacterium tuberculosis Is a Nucleoid-Associated Protein," *PLoS Pathog.*, vol. 8, no. 3, p. e1002621, Mar. 2012, doi: 10.1371/journal.ppat.1002621.

[236] A. Trauner, K. E. A. Lougheed, M. H. Bennett, S. M. Hingley-Wilson, and H. D. Williams, "The Dormancy Regulator DosR Controls Ribosome Stability in Hypoxic Mycobacteria," *J. Biol. Chem.*, vol. 287, no. 28, pp. 24053–24063, Jul. 2012, doi: 10.1074/jbc.M112.364851.

[237] K. Raman, A. G. Bhat, and N. Chandra, "A systems perspective of host–pathogen interactions: predicting disease outcome in tuberculosis," *Mol. BioSyst.*, vol. 6, no. 3, pp. 516–530, Mar. 2010, doi: 10.1039/B912129C.

[238] S. Marino, M. El-Kebir, and D. Kirschner, "A hybrid multi-compartment model of granuloma formation and T cell priming in Tuberculosis," *J. Theor. Biol.*, vol. 280, no. 1, pp. 50–62, Jul. 2011, doi: 10.1016/j.jtbi.2011.03.022.

[239] R. L. Leistikow, R. a Morton, I. L. Bartek, I. Frimpong, K. Wagner, and M. I. Voskuil, "The Mycobacterium tuberculosis DosR Regulon Assists in Metabolic Homeostasis and Enables Rapid Recovery from Nonrespiring Dormancy," *J. Bacteriol.*, vol. 192, no.

6, pp. 1662–1670, Mar. 2010, doi: 10.1128/JB.00926-09.

[240] U. S. Gautam, S. Chauhan, and J. S. Tyagi, "Determinants Outside the DevR C-Terminal Domain Are Essential for Cooperativity and Robust Activation of Dormancy Genes in Mycobacterium tuberculosis," *PLoS One*, vol. 6, no. 1, p. e16500, Jan. 2011, doi: 10.1371/journal.pone.0016500.

[241] S. Chauhan and J. S. Tyagi, "Cooperative Binding of Phosphorylated DevR to Upstream Sites Is Necessary and Sufficient for Activation of the Rv3134c- devRS Operon in Mycobacterium tuberculosis : Implication in the Induction of DevR Target Genes," *J. Bacteriol.*, vol. 190, no. 12, pp. 4301–4312, Jun. 2008, doi: 10.1128/JB.01308-07.

[242] S. Chauhan, D. Sharma, A. Singh, A. Surolia, and J. S. Tyagi, "Comprehensive insights into Mycobacterium tuberculosis DevR (DosR) regulon activation switch," *Nucleic Acids Res.*, vol. 39, no. 17, pp. 7400–7414, Sep. 2011, doi: 10.1093/nar/gkr375.

[243] A. Kumar, J. C. Toledo, R. P. Patel, J. R. Lancaster, and A. J. C. Steyn, "Mycobacterium tuberculosis DosS is a redox sensor and DosT is a hypoxia sensor," *Proc. Natl. Acad. Sci.*, vol. 104, no. 28, pp. 11568–11573, Jul. 2007, doi: 10.1073/pnas.0705054104.

[244] R. W. Honaker, R. L. Leistikow, I. L. Bartek, and M. I. Voskuil, "Unique Roles of DosT and DosS in DosR Regulon Induction and Mycobacterium tuberculosis Dormancy," *Infect. Immun.*, vol. 77, no. 8, pp. 3258–3263, Aug. 2009, doi: 10.1128/IAI.01449-08.

[245] K. Kaur, P. Kumari, S. Sharma, S. Sehgal, and J. S. Tyagi, "DevS/DosS sensor is bifunctional and its phosphatase activity precludes aerobic DevR/DosR regulon expression in Mycobacterium tuberculosis," *FEBS J.*, vol. 283, no. 15, pp. 2949–2962, Aug. 2016, doi: 10.1111/febs.13787.

[246] R. W. Honaker, R. K. Dhiman, P. Narayanasamy, D. C. Crick, and M. I. Voskuil, "DosS Responds to a Reduced Electron Transport System To Induce the Mycobacterium tuberculosis DosR Regulon," *J. Bacteriol.*, vol. 192, no. 24, pp. 6447–6455, Dec. 2010, doi: 10.1128/JB.00978-10.

[247] A. Kumar *et al.*, "Heme Oxygenase-1-derived Carbon Monoxide Induces the Mycobacterium tuberculosis Dormancy Regulon," *J. Biol. Chem.*, vol. 283, no. 26, pp. 18032–18039, Jun. 2008, doi: 10.1074/jbc.M802274200.

[248] S. Silva-Gomes, R. Appelberg, R. Larsen, M. P. Soares, and M. S. Gomes, "Heme Catabolism by Heme Oxygenase-1 Confers Host Resistance to Mycobacterium Infection," *Infect. Immun.*, vol. 81, no. 7, pp. 2536–2545, Jul. 2013, doi: 10.1128/IAI.00251-13.

[249] R. D. Bunker *et al.*, "A functional role of Rv1738 in Mycobacterium tuberculosis persistence suggested by racemic protein crystallography," *Proc. Natl. Acad. Sci.*, vol. 112, no. 14, pp. 4310–4315, Apr. 2015, doi: 10.1073/pnas.1422387112.

[250] A. E. Maris *et al.*, "Dimerization allows DNA target site recognition by the NarL response regulator," *Nat. Struct. Biol.*, vol. 9, no. 10, pp. 771–778, Oct. 2002, doi: 10.1038/nsb845.

[251] H.-N. Lee, K.-E. Jung, I.-J. Ko, H. S. Baik, and J.-I. Oh, "Protein-protein interactions between histidine kinases and response regulators of Mycobacterium tuberculosis H37Rv," *J. Microbiol.*, vol. 50, no. 2, pp. 270–277, Apr. 2012, doi: 10.1007/s12275-012-2050-4.

[252] J.-Y. Jung *et al.*, "The Intracellular Environment of Human Macrophages That Produce Nitric Oxide Promotes Growth of Mycobacteria," *Infect. Immun.*, vol. 81, no.

9, pp. 3198–3209, Sep. 2013, doi: 10.1128/IAI.00611-13.

[253] S. Nambu, T. Matsui, C. W. Goulding, S. Takahashi, and M. Ikeda-Saito, "A New Way to Degrade Heme," *J. Biol. Chem.*, vol. 288, no. 14, pp. 10101–10109, Apr. 2013, doi: 10.1074/jbc.M112.448399.

[254] S. Arya *et al.*, "Truncated Hemoglobin, HbN, Is Post-translationally Modified in Mycobacterium tuberculosis and Modulates Host-Pathogen Interactions during Intracellular Infection," *J. Biol. Chem.*, vol. 288, no. 41, pp. 29987–29999, Oct. 2013, doi: 10.1074/jbc.M113.507301.

[255] M. V. Tullius *et al.*, "Discovery and characterization of a unique mycobacterial heme acquisition system," *Proc. Natl. Acad. Sci.*, vol. 108, no. 12, pp. 5051–5056, Mar. 2011, doi: 10.1073/pnas.1009516108.

[256] S. V. Joseph, G. K. Madhavilatha, R. A. Kumar, and S. Mundayoor, "Comparative Analysis of Mycobacterial Truncated Hemoglobin Promoters and the groEL2 Promoter in Free-Living and Intracellular Mycobacteria," *Appl. Environ. Microbiol.*, vol. 78, no. 18, pp. 6499–6506, Sep. 2012, doi: 10.1128/AEM.01984-12.

[257] D. Sethi *et al.*, "Lipoprotein LprI of Mycobacterium tuberculosis Acts as a Lysozyme Inhibitor," *J. Biol. Chem.*, vol. 291, no. 6, pp. 2938–2953, Feb. 2016, doi: 10.1074/jbc.M115.662593.

[258] B. Phetsuksiri *et al.*, "Antimycobacterial Activities of Isoxyl and New Derivatives through the Inhibition of Mycolic Acid Synthesis," *Antimicrob. Agents Chemother.*, vol. 43, no. 5, pp. 1042–1051, May 1999, doi: 10.1128/AAC.43.5.1042.

[259] G. Hall, T. D. Bradshaw, C. A. Laughton, M. F. Stevens, and J. Emsley, "Structure of Mycobacterium tuberculosis thioredoxin in complex with quinol inhibitor PMX464," *Protein Sci.*, vol. 20, no. 1, pp. 210–215, Jan. 2011, doi: 10.1002/pro.533.

[260] P. Ascenzi *et al.*, "Isoniazid Inhibits the Heme-Based Reactivity of Mycobacterium tuberculosis Truncated Hemoglobin N," *PLoS One*, vol. 8, no. 8, p. e69762, Aug. 2013, doi: 10.1371/journal.pone.0069762.

[261] M. P. Tan *et al.*, "Nitrate Respiration Protects Hypoxic Mycobacterium tuberculosis Against Acid- and Reactive Nitrogen Species Stresses," *PLoS One*, vol. 5, no. 10, p. e13356, Oct. 2010, doi: 10.1371/journal.pone.0013356.

[262] A. Khan and D. Sarkar, "Nitrate reduction pathways in mycobacteria and their implications during latency," *Microbiology*, vol. 158, no. 2, pp. 301–307, Feb. 2012, doi: 10.1099/mic.0.054759-0.

[263] T. R. Rustad, M. I. Harrell, R. Liao, and D. R. Sherman, "The Enduring Hypoxic Response of Mycobacterium tuberculosis," *PLoS One*, vol. 3, no. 1, p. e1502, Jan. 2008, doi: 10.1371/journal.pone.0001502.

[264] A. V. Veatch and D. Kaushal, "Opening Pandora's Box: Mechanisms of Mycobacterium tuberculosis Resuscitation," *Trends Microbiol.*, vol. 26, no. 2, pp. 145–157, Feb. 2018, doi: 10.1016/j.tim.2017.08.001.

[265] Y. Zhang *et al.*, "CRP Acts as a Transcriptional Repressor of the YPO1635-phoPQ-YPO1632 Operon in Yersinia pestis," *Curr. Microbiol.*, vol. 70, no. 3, pp. 398–403, Mar. 2015, doi: 10.1007/s00284-014-0736-z.

[266] R. K. Gupta, B. S. Srivastava, and R. Srivastava, "Comparative expression analysis of rpf-like genes of Mycobacterium tuberculosis H37Rv under different physiological stress and growth conditions," *Microbiology*, vol. 156, no. 9, pp. 2714–2722, Sep. 2010, doi: 10.1099/mic.0.037622-0.

[267] G. Bai, L. A. McCue, and K. A. McDonough, “Characterization of Mycobacterium tuberculosis Rv3676 (CRP Mt), a Cyclic AMP Receptor Protein-Like DNA Binding Protein,” *J. Bacteriol.*, vol. 187, no. 22, pp. 7795–7804, Nov. 2005, doi: 10.1128/JB.187.22.7795-7804.2005.

[268] L. Rickman *et al.*, “A member of the cAMP receptor protein family of transcription regulators in Mycobacterium tuberculosis is required for virulence in mice and controls transcription of the rpfA gene coding for a resuscitation promoting factor,” *Mol. Microbiol.*, vol. 56, no. 5, pp. 1274–1286, Mar. 2005, doi: 10.1111/j.1365-2958.2005.04609.x.

[269] D. J. Bretl *et al.*, “MprA and DosR Coregulate a Mycobacterium tuberculosis Virulence Operon Encoding Rv1813c and Rv1812c,” *Infect. Immun.*, vol. 80, no. 9, pp. 3018–3033, Sep. 2012, doi: 10.1128/IAI.00520-12.

[270] X. Pang, G. Cao, P. F. Neuenschwander, S. E. Haydel, G. Hou, and S. T. Howard, “The β-propeller gene Rv1057 of Mycobacterium tuberculosis has a complex promoter directly regulated by both the MprAB and TrcRS two-component systems,” *Tuberculosis*, vol. 91, pp. S142–S149, Dec. 2011, doi: 10.1016/j.tube.2011.10.024.

[271] X. Pang *et al.*, “Evidence for complex interactions of stress-associated regulons in an mprAB deletion mutant of Mycobacterium tuberculosis,” *Microbiology*, vol. 153, no. 4, pp. 1229–1242, Apr. 2007, doi: 10.1099/mic.0.29281-0.

[272] S. V. Jamwal, P. Mehrotra, A. Singh, Z. Siddiqui, A. Basu, and K. V. S. Rao, “Mycobacterial escape from macrophage phagosomes to the cytoplasm represents an alternate adaptation mechanism,” *Sci. Rep.*, vol. 6, no. 1, p. 23089, Sep. 2016, doi: 10.1038/srep23089.

[273] Y. Zhang *et al.*, “Autoregulation of PhoP/PhoQ and Positive Regulation of the Cyclic AMP Receptor Protein-Cyclic AMP Complex by PhoP in Yersinia pestis,” *J. Bacteriol.*, vol. 195, no. 5, pp. 1022–1030, Mar. 2013, doi: 10.1128/JB.01530-12.

[274] M. R. Jofré, L. M. Rodríguez, N. A. Villagra, A. A. Hidalgo, G. C. Mora, and J. A. Fuentes, “RpoS integrates CRP, Fis, and PhoP signaling pathways to control Salmonella Typhi hlyE expression,” *BMC Microbiol.*, vol. 14, no. 1, p. 139, Dec. 2014, doi: 10.1186/1471-2180-14-139.

[275] M. R. W. Brown and A. Kornberg, “Inorganic polyphosphate in the origin and survival of species,” *Proc. Natl. Acad. Sci.*, vol. 101, no. 46, pp. 16085–16087, Nov. 2004, doi: 10.1073/pnas.0406909101.

[276] Y. Chuang, N. K. Dutta, C. Hung, T.-C. Wu, H. Rubin, and P. C. Karakousis, “Stringent Response Factors PPX1 and PPK2 Play an Important Role in Mycobacterium tuberculosis Metabolism, Biofilm Formation, and Sensitivity to Isoniazid In Vivo,” *Antimicrob. Agents Chemother.*, vol. 60, no. 11, pp. 6460–6470, Nov. 2016, doi: 10.1128/AAC.01139-16.

[277] S. Sanyal, S. K. Banerjee, R. Banerjee, J. Mukhopadhyay, and M. Kundu, “Polyphosphate kinase 1, a central node in the stress response network of Mycobacterium tuberculosis, connects the two-component systems MprAB and SenX3–RegX3 and the extracytoplasmic function sigma factor, sigma E,” *Microbiology*, vol. 159, no. Pt_10, pp. 2074–2086, Oct. 2013, doi: 10.1099/mic.0.068452-0.

[278] R. Manganelli and R. Provvedi, “An integrated regulatory network including two positive feedback loops to modulate the activity of σ E in mycobacteria,” *Mol. Microbiol.*, vol. 75, no. 3, pp. 538–542, Feb. 2010, doi: 10.1111/j.1365-2958.2009.07009.x.

[279] J. Troudt *et al.*, "Mycobacterium tuberculosis sigE mutant ST28 used as a vaccine induces protective immunity in the guinea pig model," *Tuberculosis*, vol. 106, pp. 99–105, Sep. 2017, doi: 10.1016/j.tube.2017.07.009.

[280] V. Chelliah *et al.*, "BioModels: ten-year anniversary," *Nucleic Acids Res.*, vol. 43, no. D1, pp. D542–D548, Jan. 2015, doi: 10.1093/nar/gku1181.

[281] J. P. McCutcheon and N. A. Moran, "Extreme genome reduction in symbiotic bacteria," *Nat. Rev. Microbiol.*, vol. 10, no. 1, pp. 13–26, Jan. 2012, doi: 10.1038/nrmicro2670.

[282] J. Gatfield and J. Pieters, "Essential Role for Cholesterol in Entry of Mycobacteria into Macrophages," *Science (80-.).*, vol. 288, no. 5471, pp. 1647–1651, Jun. 2000, doi: 10.1126/science.288.5471.1647.

[283] F. J. Veyrier, A. Dufort, and M. A. Behr, "The rise and fall of the Mycobacterium tuberculosis genome," *Trends Microbiol.*, vol. 19, no. 4, pp. 156–161, Apr. 2011, doi: 10.1016/j.tim.2010.12.008.

[284] P. Supply *et al.*, "Genomic analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of Mycobacterium tuberculosis," *Nat. Genet.*, vol. 45, no. 2, pp. 172–179, Feb. 2013, doi: 10.1038/ng.2517.

[285] Q. Chai, Y. Zhang, and C. H. Liu, "Mycobacterium tuberculosis: An Adaptable Pathogen Associated With Multiple Human Diseases," *Front. Cell. Infect. Microbiol.*, vol. 8, no. MAY, pp. 1–15, May 2018, doi: 10.3389/fcimb.2018.00158.

[286] R. Pal, M. K. Bisht, and S. Mukhopadhyay, "Secretory proteins of Mycobacterium tuberculosis and their roles in modulation of host immune responses: focus on therapeutic targets," *FEBS J.*, vol. 289, no. 14, pp. 4146–4171, Jul. 2022, doi: 10.1111/febs.16369.

[287] F. Veyrier, D. Pletzer, C. Turenne, and M. A. Behr, "Phylogenetic detection of horizontal gene transfer during the step-wise genesis of Mycobacterium tuberculosis," *BMC Evol. Biol.*, vol. 9, no. 1, p. 196, Dec. 2009, doi: 10.1186/1471-2148-9-196.

[288] C. G. Korea *et al.*, "Staphylococcal Esx Proteins Modulate Apoptosis and Release of Intracellular Staphylococcus aureus during Infection in Epithelial Cells," *Infect. Immun.*, vol. 82, no. 10, pp. 4144–4153, Oct. 2014, doi: 10.1128/IAI.01576-14.

[289] L. Lai *et al.*, "Streptococcus suis serotype 9 strain GZ0565 contains a type VII secretion system putative substrate EsxA that contributes to bacterial virulence and a vanZ- like gene that confers resistance to teicoplanin and dalbavancin in Streptococcus agalactiae," *Vet. Microbiol.*, vol. 205, pp. 26–33, Jun. 2017, doi: 10.1016/j.vetmic.2017.04.030.

[290] C. Vullo, "Inactivation of selected genes associated to ESX loci for the development of new antituberculosis vaccine candidates," *Thesis etd-12212021-134838*, 2022, [Online]. Available: https://etd.adm.unipi.it/t/etd-12212021-134838/

[291] D. Young, J. Stark, and D. Kirschner, "Systems biology of persistent infection: tuberculosis as a case study," *Nat. Rev. Microbiol.*, vol. 6, no. 7, pp. 520–528, Jul. 2008, doi: 10.1038/nrmicro1919.

[292] R. De Smet and K. Marchal, "Advantages and limitations of current network inference methods," *Nat. Rev. Microbiol.*, vol. 8, no. 10, pp. 717–729, Oct. 2010, doi: 10.1038/nrmicro2419.

[293] D. F. T. T. Veiga, B. Dutta, and G. Balázsi, "Network inference and network response identification: moving genome-scale data to the next level of biological discovery,"

Mol. BioSyst., vol. 6, no. 3, pp. 469–480, 2010, doi: 10.1039/B916989J.

[294] M. Van Iersel, "Uses for Pathways PathVisio WikiPathways," 2010.

[295] A. M. Abdallah *et al.*, "Type VII secretion — mycobacteria show the way," *Nat. Rev. Microbiol.*, vol. 5, no. 11, pp. 883–891, Nov. 2007, doi: 10.1038/nrmicro1773.

[296] H.-D. Park *et al.*, "Rv3133c/dosR is a transcription factor that mediates the hypoxic response of Mycobacterium tuberculosis," *Mol. Microbiol.*, vol. 48, no. 3, pp. 833–843, Apr. 2003, doi: 10.1046/j.1365-2958.2003.03474.x.

[297] S. Mehra *et al.*, "The DosR Regulon Modulates Adaptive Immunity and Is Essential for Mycobacterium tuberculosis Persistence," *Am. J. Respir. Crit. Care Med.*, vol. 191, no. 10, pp. 1185–1196, May 2015, doi: 10.1164/rccm.201408-1502OC.

[298] T. L. Bailey, "Discovering Novel Sequence Motifs with <scp>MEME</scp>," *Curr. Protoc. Bioinforma.*, vol. 00, no. 1, pp. 28–36, Jan. 2003, doi: 10.1002/0471250953.bi0204s00.

[299] A. A. Margolin *et al.*, "ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context," *BMC Bioinformatics*, vol. 7, no. S1, p. S7, Mar. 2006, doi: 10.1186/1471-2105-7-S1-S7.

[300] J. M. Lew, A. Kapopoulou, L. M. Jones, and S. T. Cole, "Tuberculist, Release 27 - March 2013," 2013. http://www.http//tuberculist.epfl.ch

[301] P. D. Karp *et al.*, "The BioCyc collection of microbial genomes and metabolic pathways," *Brief. Bioinform.*, vol. 20, no. 4, pp. 1085–1093, Jul. 2019, doi: 10.1093/bib/bbx085.

[302] D. Szklarczyk *et al.*, "The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D362–D368, Jan. 2017, doi: 10.1093/nar/gkw937.

[303] A. Banerjee *et al.*, "A Universal Stress Protein (USP) in Mycobacteria Binds cAMP," *J. Biol. Chem.*, vol. 290, no. 20, pp. 12731–12743, May 2015, doi: 10.1074/jbc.M115.644856.

[304] J. E. Drumm *et al.*, "Mycobacterium tuberculosis Universal Stress Protein Rv2623 Regulates Bacillary Growth by ATP-Binding: Requirement for Establishing Chronic Persistent Infection," *PLoS Pathog.*, vol. 5, no. 5, p. e1000460, May 2009, doi: 10.1371/journal.ppat.1000460.

[305] L. N. Glass *et al.*, "Mycobacterium tuberculosis universal stress protein Rv2623 interacts with the putative ATP binding cassette (ABC) transporter Rv1747 to regulate mycobacterial growth," *PLOS Pathog.*, vol. 13, no. 7, p. e1006515, Jul. 2017, doi: 10.1371/journal.ppat.1006515.

[306] C. R. E. McEvoy *et al.*, "Comparative Analysis of Mycobacterium tuberculosis pe and ppe Genes Reveals High Sequence Variation and an Apparent Absence of Selective Constraints," *PLoS One*, vol. 7, no. 4, p. e30593, Apr. 2012, doi: 10.1371/journal.pone.0030593.

[307] P. Chaiyachat *et al.*, "Whole-genome analysis of drug-resistant Mycobacterium tuberculosis reveals novel mutations associated with fluoroquinolone resistance," *Int. J. Antimicrob. Agents*, vol. 58, no. 3, p. 106385, Sep. 2021, doi: 10.1016/j.ijantimicag.2021.106385.

[308] S. J. Modlin *et al.*, "Drivers and sites of diversity in the DNA adenine methylomes of 93 Mycobacterium tuberculosis complex clinical isolates," *Elife*, vol. 9, pp. 1–33, Oct. 2020, doi: 10.7554/eLife.58542.

[309] A. Kumar *et al.*, "Mycobacterium tuberculosis DosR Regulon Gene Rv0079 Encodes a Putative, 'Dormancy Associated Translation Inhibitor (DATIN),'" *PLoS One*, vol. 7, no. 6, p. e38709, Jun. 2012, doi: 10.1371/journal.pone.0038709.

[310] N. D. Fernandes, Q.-L. Wu, D. Kong, X. Puyang, S. Garg, and R. N. Husson, "A Mycobacterial Extracytoplasmic Sigma Factor Involved in Survival following Heat Shock and Oxidative Stress," *J. Bacteriol.*, vol. 181, no. 19, pp. 6222–6222, Oct. 1999, doi: 10.1128/JB.181.19.6222-6222.1999.

[311] M. J. White, H. He, R. M. Penoske, S. S. Twining, and T. C. Zahrt, "PepD Participates in the Mycobacterial Stress Response Mediated through MprAB and SigE," *J. Bacteriol.*, vol. 192, no. 6, pp. 1498–1510, Mar. 2010, doi: 10.1128/JB.01167-09.

[312] S. Casonato, R. Provvedi, E. Dainese, G. Palù, and R. Manganelli, "Mycobacterium tuberculosis Requires the ECF Sigma Factor SigE to Arrest Phagosome Maturation," *PLoS One*, vol. 9, no. 9, p. e108893, Sep. 2014, doi: 10.1371/journal.pone.0108893.

[313] R. Hernandez-Pando *et al.*, "Construction and Characterization of the Mycobacterium tuberculosis sigE fadD26 Unmarked Double Mutant as a Vaccine Candidate," *Infect. Immun.*, vol. 88, no. 1, pp. 1–11, Dec. 2019, doi: 10.1128/IAI.00496-19.

[314] S. Chauhan and J. S. Tyagi, "Interaction of DevR with Multiple Binding Sites Synergistically Activates Divergent Transcription of narK2 -Rv1738 Genes in Mycobacterium tuberculosis," *J. Bacteriol.*, vol. 190, no. 15, pp. 5394–5403, Aug. 2008, doi: 10.1128/JB.00488-08.

[315] H. He, R. Hovey, J. Kane, V. Singh, and T. C. Zahrt, "Correction for He et al., 'MprAB Is a Stress-Responsive Two-Component System That Directly Regulates Expression of Sigma Factors SigB and SigE in Mycobacterium tuberculosis,'" *J. Bacteriol.*, vol. 202, no. 20, pp. 2134–2143, Sep. 2020, doi: 10.1128/JB.00443-20.

[316] S. Barik, K. Sureka, P. Mukherjee, J. Basu, and M. Kundu, "RseA, the SigE specific anti-sigma factor of Mycobacterium tuberculosis , is inactivated by phosphorylation-dependent ClpC1P2 proteolysis," *Mol. Microbiol.*, vol. 75, no. 3, pp. 592–606, Feb. 2010, doi: 10.1111/j.1365-2958.2009.07008.x.

[317] M. R. Graham *et al.*, "Virulence control in group A Streptococcus by a two-component gene regulatory system: Global expression profiling and in vivo infection modeling," *Proc. Natl. Acad. Sci.*, vol. 99, no. 21, pp. 13855–13860, Oct. 2002, doi: 10.1073/pnas.202353699.

[318] B. Krismer, C. Weidenmaier, A. Zipperer, and A. Peschel, "The commensal lifestyle of Staphylococcus aureus and its interactions with the nasal microbiota," *Nat. Rev. Microbiol.*, vol. 15, no. 11, pp. 675–687, Nov. 2017, doi: 10.1038/nrmicro.2017.104.

[319] B. Henriques-Normark and S. Normark, "Commensal pathogens, with a focus on Streptococcus pneumoniae, and interactions with the human host," *Exp. Cell Res.*, vol. 316, no. 8, pp. 1408–1414, May 2010, doi: 10.1016/j.yexcr.2010.03.003.

[320] E. Török and N. Day, "Staphylococcal and streptococcal infections," *Medicine (Baltimore).*, vol. 33, no. 5, pp. 97–100, May 2005, doi: 10.1383/medc.33.5.97.64964.

[321] N. Patenge, R. Pappesch, A. Khani, and B. Kreikemeyer, "Genome-wide analyses of small non-coding RNAs in streptococci," *Front. Genet.*, vol. 06, no. MAY, pp. 1–13, May 2015, doi: 10.3389/fgene.2015.00189.

[322] H. Suzuki, T. Lefébure, P. P. Bitar, and M. J. Stanhope, "Comparative genomic analysis of the genus Staphylococcus including Staphylococcus aureus and its newly described sister species Staphylococcus simiae," *BMC Genomics*, vol. 13, no. 1, p. 38, Dec. 2012, doi: 10.1186/1471-2164-13-38.

[323] E. Saccenti, D. Nieuwenhuijse, J. J. Koehorst, V. A. P. Martins dos Santos, and P. J. Schaap, "Assessing the Metabolic Diversity of Streptococcus from a Protein Domain Point of View," *PLoS One*, vol. 10, no. 9, p. e0137908, Sep. 2015, doi: 10.1371/journal.pone.0137908.

[324] E. Bosi, J. M. Monk, R. K. Aziz, M. Fondi, V. Nizet, and B. Ø. Palsson, "Comparative genome-scale modelling of Staphylococcus aureus strains identifies strain-specific metabolic capabilities linked to pathogenicity," *Proc. Natl. Acad. Sci.*, vol. 113, no. 26, pp. E3801–E3809, Jun. 2016, doi: 10.1073/pnas.1523199113.

[325] J. J. Koehorst *et al.*, "Comparison of 432 Pseudomonas strains through integration of genomic, functional, metabolic and expression data," *Sci. Rep.*, vol. 6, no. 1, p. 38699, Dec. 2016, doi: 10.1038/srep38699.

[326] J. J. Koehorst, E. Saccenti, P. J. Schaap, V. A. P. Martins dos Santos, and M. Suarez-Diez, "Protein domain architectures provide a fast, efficient and scalable alternative to sequence-based methods for comparative functional genomics," *F1000Research*, vol. 5, no. 0, p. 1987, Jun. 2017, doi: 10.12688/f1000research.9416.3.

[327] L. Rouli, V. Merhej, P.-E. Fournier, and D. Raoult, "The bacterial pangenome as a new tool for analysing pathogenic bacteria," *New Microbes New Infect.*, vol. 7, pp. 72–85, Sep. 2015, doi: 10.1016/j.nmni.2015.06.005.

[328] S. Fuchs *et al.*, "Aureo Wiki-The repository of the Staphylococcus aureus research and annotation community," *Int. J. Med. Microbiol.*, vol. 308, no. 6, pp. 558–568, Aug. 2018, doi: 10.1016/j.ijmm.2017.11.011.

[329] X.-Y. Gao, X.-Y. Zhi, H.-W. Li, H.-P. Klenk, and W.-J. Li, "Comparative Genomics of the Bacterial Genus Streptococcus Illuminates Evolutionary Implications of Species Groups," *PLoS One*, vol. 9, no. 6, p. e101229, Jun. 2014, doi: 10.1371/journal.pone.0101229.

[330] L. Snipen, T. Almøy, and D. W. Ussery, "Microbial comparative pan-genomics using binomial mixture models," *BMC Genomics*, vol. 10, no. 1, p. 385, 2009, doi: 10.1186/1471-2164-10-385.

[331] H. Tettelin, D. Riley, C. Cattuto, and D. Medini, "Comparative genomics: the bacterial pan-genome," *Curr. Opin. Microbiol.*, vol. 11, no. 5, pp. 472–477, Oct. 2008, doi: 10.1016/j.mib.2008.09.006.

[332] M. Kuroda *et al.*, "Whole genome sequencing of meticillin-resistant Staphylococcus aureus," *Lancet*, vol. 357, no. 9264, pp. 1225–1240, Apr. 2001, doi: 10.1016/S0140-6736(00)04403-2.

[333] S. A. Matyi *et al.*, "Isolation and characterization of Staphylococcus aureus strains from a Paso del Norte dairy," *J. Dairy Sci.*, vol. 96, no. 6, pp. 3535–3542, Jun. 2013, doi: 10.3168/jds.2013-6590.

[334] F. T., "Staphylococcus, Chapter 12 in Microbiology. 4th edition," in *Microbiology. 4th edition*, B. S, Ed. Galveston (TX): University of Texas Medical Branch at Galveston, 1996, p. Chapter 12. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK8448/996.

[335] Y. Li, B. Cao, Y. Zhang, J. Zhou, B. Yang, and L. Wang, "Complete Genome Sequence of Staphylococcus aureus T0131, an ST239-MRSA-SCC mec Type III Clone Isolated in China," *J. Bacteriol.*, vol. 193, no. 13, pp. 3411–3412, Jul. 2011, doi: 10.1128/JB.05135-11.

[336] P. Sass *et al.*, "Genome Sequence of Staphylococcus aureus VC40, a Vancomycin- and Daptomycin-Resistant Strain, To Study the Genetics of Development of Resistance to

Currently Applied Last-Resort Antibiotics," *J. Bacteriol.*, vol. 194, no. 8, pp. 2107–2108, Apr. 2012, doi: 10.1128/JB.06631-11.

[337] T. Baba, T. Bae, O. Schneewind, F. Takeuchi, and K. Hiramatsu, "Genome Sequence of Staphylococcus aureus Strain Newman and Comparative Analysis of Staphylococcal Genomes: Polymorphism and Evolution of Two Major Pathogenicity Islands," *J. Bacteriol.*, vol. 190, no. 1, pp. 300–310, Jan. 2008, doi: 10.1128/JB.01000-07.

[338] A. E. Zautner *et al.*, "Intracellular Persisting Staphylococcus aureus Is the Major Pathogen in Recurrent Tonsillitis," *PLoS One*, vol. 5, no. 3, p. e9452, Mar. 2010, doi: 10.1371/journal.pone.0009452.

[339] K. Surmann *et al.*, "Analysis of Staphylococcus aureus proteins secreted inside infected human epithelial cells," *Int. J. Med. Microbiol.*, vol. 308, no. 6, pp. 664–674, Aug. 2018, doi: 10.1016/j.ijmm.2018.06.002.

[340] P. Sendi and R. A. Proctor, "Staphylococcus aureus as an intracellular pathogen: the role of small colony variants," *Trends Microbiol.*, vol. 17, no. 2, pp. 54–58, Feb. 2009, doi: 10.1016/j.tim.2008.11.004.

[341] B. Kreikemeyer, K. S. McIver, and A. Podbielski, "Virulence factor regulation and regulatory networks in Streptococcus pyogenes and their impact on pathogen–host interactions," *Trends Microbiol.*, vol. 11, no. 5, pp. 224–232, May 2003, doi: 10.1016/S0966-842X(03)00098-2.

[342] T. G. Loof, C. Deicke, and E. Medina, "The role of coagulation/fibrinolysis during Streptococcus pyogenes infection," *Front. Cell. Infect. Microbiol.*, vol. 4, no. September, pp. 1–8, Sep. 2014, doi: 10.3389/fcimb.2014.00128.

[343] A. Jenkins *et al.*, "Differential Expression and Roles of Staphylococcus aureus Virulence Determinants during Colonization and Disease," *MBio*, vol. 6, no. 1, pp. 1–10, Feb. 2015, doi: 10.1128/mBio.02272-14.

[344] T. Acharya and T. Acharya, "Virulence factors of Streptococcus pyogenes and their roles," *Microbe online*, 2016. https://microbeonline.com/virulence-factors-streptococcus-pyogenes-roles/ (accessed Dec. 17, 2018).

[345] S. Z. Alborzi, M. Devignes, and D. W. Ritchie, "Associating Gene Ontology Terms with Pfam Protein Domains," in *Hal-Inria*, 2017, pp. 127–138. doi: 10.1007/978-3-319-56154-7_13.

[346] F. Askarian *et al.*, "The interaction between Staphylococcus aureus SdrD and desmoglein 1 is important for adhesion to host cells," *Sci. Rep.*, vol. 6, no. 1, p. 22134, Apr. 2016, doi: 10.1038/srep22134.

[347] N. H. R. Eriksen, F. Espersen, V. T. Rosdahl, and K. Jensen, "Carriage of Staphylococcus aureus among 104 healthy persons during a 19-month period," *Epidemiol. Infect.*, vol. 115, no. 1, pp. 51–60, Aug. 1995, doi: 10.1017/S0950268800058118.

[348] S. J. Hermans, H. M. Baker, R. P. Sequeira, R. J. Langley, E. N. Baker, and J. D. Fraser, "Structural and Functional Properties of Staphylococcal Superantigen-Like Protein 4," *Infect. Immun.*, vol. 80, no. 11, pp. 4004–4013, Nov. 2012, doi: 10.1128/IAI.00764-12.

[349] V. Stemberk *et al.*, "Evidence for Steric Regulation of Fibrinogen Binding to Staphylococcus aureus Fibronectin-binding Protein A (FnBPA)," *J. Biol. Chem.*, vol. 289, no. 18, pp. 12842–12851, May 2014, doi: 10.1074/jbc.M113.543546.

[350] B. C. Fries and A. K. Varshney, "Bacterial Toxins—Staphylococcal Enterotoxin B,"

Microbiol. Spectr., vol. 1, no. 2, pp. 1–12, Dec. 2013, doi: 10.1128/microbiolspec.AID-0002-2012.

[351] M. Gottlieb, B. Long, and A. Koyfman, "The Evaluation and Management of Toxic Shock Syndrome in the Emergency Department: A Review of the Literature," *J. Emerg. Med.*, vol. 54, no. 6, pp. 807–814, Jun. 2018, doi: 10.1016/j.jemermed.2017.12.048.

[352] M. Otto, "Staphylococcus aureus toxins," *Curr. Opin. Microbiol.*, vol. 17, pp. 32–37, Feb. 2014, doi: 10.1016/j.mib.2013.11.004.

[353] S. L. Kolar *et al.*, "Extracellular proteases are key mediators of <scp> *S* </scp> *taphylococcus aureus* virulence via the global modulation of virulence-determinant stability," *Microbiologyopen*, vol. 2, no. 1, pp. 18–34, Feb. 2013, doi: 10.1002/mbo3.55.

[354] W.-H. Chen, G. Lu, X. Chen, X.-M. Zhao, and P. Bork, "OGEE v2: an update of the online gene essentiality database with special focus on differentially essential genes in human cancer cell lines," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D940–D944, Jan. 2017, doi: 10.1093/nar/gkw1013.

[355] J. G. Bohanek and R. Fivush, "Personal narratives, well-being, and gender in adolescence," *Cogn. Dev.*, vol. 25, no. 4, pp. 368–379, Oct. 2010, doi: 10.1016/j.cogdev.2010.08.003.

[356] M. Rohde and P. P. Cleary, *Adhesion and invasion of Streptococcus pyogenes into host cells and clinical relevance of intracellular streptococci*. 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK333420/

[357] U. Mäder *et al.*, "Staphylococcus aureus Transcriptome Architecture: From Laboratory to Infection-Mimicking Conditions," *PLOS Genet.*, vol. 12, no. 4, p. e1005962, Apr. 2016, doi: 10.1371/journal.pgen.1005962.

[358] R. R. Chaudhuri *et al.*, "Comprehensive identification of essential Staphylococcus aureus genes using Transposon-Mediated Differential Hybridisation (TMDH)," *BMC Genomics*, vol. 10, no. 1, p. 291, Dec. 2009, doi: 10.1186/1471-2164-10-291.

[359] R. A. Forsyth *et al.*, "A genome-wide strategy for the identification of essential genes in Staphylococcus aureus," *Mol. Microbiol.*, vol. 43, no. 6, pp. 1387–1400, Mar. 2002, doi: 10.1046/j.1365-2958.2002.02832.x.

[360] B. Henderson, S. Nair, J. Pallas, and M. A. Williams, "Fibronectin: a multidomain host adhesin targeted by bacterial fibronectin-binding proteins," *FEMS Microbiol. Rev.*, vol. 35, no. 1, pp. 147–200, Jan. 2011, doi: 10.1111/j.1574-6976.2010.00243.x.

[361] A. Jeng, V. Sakota, Z. Li, V. Datta, B. Beall, and V. Nizet, "Molecular Genetic Analysis of a Group A Streptococcus Operon Encoding Serum Opacity Factor and a Novel Fibronectin-Binding Protein, SfbX," *J. Bacteriol.*, vol. 185, no. 4, pp. 1208–1217, Feb. 2003, doi: 10.1128/JB.185.4.1208-1217.2003.

[362] A. M. Edwards, J. R. Potts, E. Josefsson, and R. C. Massey, "Staphylococcus aureus Host Cell Invasion and Virulence in Sepsis Is Facilitated by the Multiple Repeats within FnBPA," *PLoS Pathog.*, vol. 6, no. 6, p. e1000964, Jun. 2010, doi: 10.1371/journal.ppat.1000964.

[363] A. M. Timmer *et al.*, "Serum opacity factor promotes group A streptococcal epithelial cell invasion and virulence," *Mol. Microbiol.*, vol. 62, no. 1, pp. 15–25, Oct. 2006, doi: 10.1111/j.1365-2958.2006.05337.x.

[364] A. R. Spaulding, W. Salgado-Pabón, P. L. Kohler, A. R. Horswill, D. Y. M. Leung, and

P. M. Schlievert, “Staphylococcal and Streptococcal Superantigen Exotoxins,” *Clin. Microbiol. Rev.*, vol. 26, no. 3, pp. 422–447, Jul. 2013, doi: 10.1128/CMR.00104-12.

[365] K. Kulhankova *et al.*, “The Superantigen Toxic Shock Syndrome Toxin 1 Alters Human Aortic Endothelial Cell Function,” *Infect. Immun.*, vol. 86, no. 3, pp. 1–16, Mar. 2018, doi: 10.1128/IAI.00848-17.

[366] H. Stoll *et al.*, “Staphylococcal Enterotoxins Dose-Dependently Modulate the Generation of Myeloid-Derived Suppressor Cells,” *Front. Cell. Infect. Microbiol.*, vol. 8, no. September, pp. 1–15, Sep. 2018, doi: 10.3389/fcimb.2018.00321.

[367] T. Bae *et al.*, “Staphylococcus aureus virulence genes identified by bursa aurealis mutagenesis and nematode killing,” *Proc. Natl. Acad. Sci.*, vol. 101, no. 33, pp. 12312–12317, Aug. 2004, doi: 10.1073/pnas.0404728101.

[368] T. Kamminga *et al.*, “Persistence of Functional Protein Domains in Mycoplasma Species and their Role in Host Specificity and Synthetic Minimal Life,” *Front. Cell. Infect. Microbiol.*, vol. 7, p. 31, Feb. 2017, doi: 10.3389/fcimb.2017.00031.

[369] V. S. Cooper *et al.*, “Experimental Evolution In Vivo To Identify Selective Pressures during Pneumococcal Colonization,” *mSystems*, vol. 5, no. 3, pp. 1–17, Jun. 2020, doi: 10.1128/mSystems.00352-20.

[370] M. Pain, E. Hjerde, C. Klingenberg, and J. P. Cavanagh, “Comparative Genomic Analysis of Staphylococcus haemolyticus Reveals Key to Hospital Adaptation and Pathogenicity,” *Front. Microbiol.*, vol. 10, no. September, pp. 1–13, Sep. 2019, doi: 10.3389/fmicb.2019.02096.

[371] M. Anisimova, J. Bielawski, K. Dunn, and Z. Yang, “Phylogenomic analysis of natural selection pressure in Streptococcus genomes,” *BMC Evol. Biol.*, vol. 7, no. 1, p. 154, 2007, doi: 10.1186/1471-2148-7-154.

[372] D. Jamrozy *et al.*, “Evolution of mobile genetic element composition in an epidemic methicillin-resistant Staphylococcus aureus: temporal changes correlated with frequent loss and gain events,” *BMC Genomics*, vol. 18, no. 1, p. 684, Dec. 2017, doi: 10.1186/s12864-017-4065-z.

[373] A. J. McCarthy, J. A. Lindsay, and A. Loeffler, “Are all meticillin-resistant Staphylococcus aureus (MRSA) equal in all hosts? Epidemiological and genetic comparison between animal and human MRSA,” *Vet. Dermatol.*, vol. 23, no. 4, pp. 267-e54, Aug. 2012, doi: 10.1111/j.1365-3164.2012.01072.x.

[374] M. Osaki, D. Takamatsu, Y. Shimoji, and T. Sekizaki, “Characterization of Streptococcus suis Genes Encoding Proteins Homologous to Sortase of Gram-Positive Bacteria,” *J. Bacteriol.*, vol. 184, no. 4, pp. 971–982, Feb. 2002, doi: 10.1128/jb.184.4.971-982.2002.

[375] C. Kao *et al.*, “Clinical and genetic analysis of invasive and non-invasive group A streptococcal infections in central Taiwan,” *J Microbiol Immunol Infect.*, pp. 105–111, 2005.

[376] C. L. McNeilly and D. J. McMillan, “Horizontal gene transfer and recombination in Streptococcus dysgalactiae subsp. equisimilis,” *Front. Microbiol.*, vol. 5, no. DEC, pp. 1–6, Dec. 2014, doi: 10.3389/fmicb.2014.00676.

[377] P. Bork and R. F. Doolittle, “Proposed acquisition of an animal protein domain by bacteria.,” *Proc. Natl. Acad. Sci.*, vol. 89, no. 19, pp. 8990–8994, Oct. 1992, doi: 10.1073/pnas.89.19.8990.

[378] G. Goyette-Desjardins, J.-P. Auger, J. Xu, M. Segura, and M. Gottschalk,

"Streptococcus suis , an important pig pathogen and emerging zoonotic agent—an update on the worldwide distribution based on serotyping and sequence typing," *Emerg. Microbes Infect.*, vol. 3, no. 1, pp. 1–20, Jan. 2014, doi: 10.1038/emi.2014.45.

[379] A. Kerdsin *et al.*, "Genotypic Profile of Streptococcus suis Serotype 2 and Clinical Features of Infection in Humans, Thailand," *Emerg. Infect. Dis.*, vol. 17, no. 5, pp. 835–842, May 2011, doi: 10.3201/eid1705.100754.

[380] H. D. T. Nghia *et al.*, "Human Case of Streptococcus suis Serotype 16 Infection," *Emerg. Infect. Dis.*, vol. 14, no. 1, pp. 155–157, Jan. 2008, doi: 10.3201/eid1401.070534.

[381] H. F. L. Wertheim, H. D. T. Nghia, W. Taylor, and C. Schultsz, "Streptococcus suis: An Emerging Human Pathogen," *Clin. Infect. Dis.*, vol. 48, no. 5, pp. 617–625, Mar. 2009, doi: 10.1086/596763.

[382] N. Hasegawa *et al.*, "Characterization of the Pathogenicity of Streptococcus intermedius TYG1620 Isolated from a Human Brain Abscess Based on the Complete Genome Sequence with Transcriptome Analysis and Transposon Mutagenesis in a Murine Subcutaneous Abscess Model," *Infect. Immun.*, vol. 85, no. 2, pp. 1–15, Feb. 2017, doi: 10.1128/IAI.00886-16.

[383] A. G. Allen *et al.*, "Generation and Characterization of a Defined Mutant of Streptococcus suis Lacking Suilysin," *Infect. Immun.*, vol. 69, no. 4, pp. 2732–2735, Apr. 2001, doi: 10.1128/IAI.69.4.2732-2735.2001.

[384] Z. HE *et al.*, "Increased production of suilysin contributes to invasive infection of the Streptococcus suis strain 05ZYH33," *Mol. Med. Rep.*, vol. 10, no. 6, pp. 2819–2826, Dec. 2014, doi: 10.3892/mmr.2014.2586.

[385] A. Remmington and C. E. Turner, "The DNases of pathogenic Lancefield streptococci," *Microbiology*, vol. 164, no. 3, pp. 242–250, Mar. 2018, doi: 10.1099/mic.0.000612.

[386] P. Sharma *et al.*, "Role of Pilus Proteins in Adherence and Invasion of Streptococcus agalactiae to the Lung and Cervical Epithelial Cells," *J. Biol. Chem.*, vol. 288, no. 6, pp. 4023–4034, Feb. 2013, doi: 10.1074/jbc.M112.425728.

[387] J. D. Bryan and D. W. Shelver, "Streptococcus agalactiae CspA Is a Serine Protease That Inactivates Chemokines," *J. Bacteriol.*, vol. 191, no. 6, pp. 1847–1854, Mar. 2009, doi: 10.1128/JB.01124-08.

[388] European Bioinformatics Institute, "EnaBroswerTools." https://github.com/enasequence/enaBrowserTools (accessed Nov. 07, 2019).

[389] J. C. J. van Dam, J. J. Koehorst, J. O. Vik, V. A. P. Martins dos Santos, P. J. Schaap, and M. Suarez-Diez, "The Empusa code generator and its application to GBOL, an extendable ontology for genome annotation," *Sci. Data*, vol. 6, no. 1, p. 254, Dec. 2019, doi: 10.1038/s41597-019-0263-7.

[390] D. Hyatt, G.-L. Chen, P. F. LoCascio, M. L. Land, F. W. Larimer, and L. J. Hauser, "Prodigal: prokaryotic gene recognition and translation initiation site identification," *BMC Bioinformatics*, vol. 11, no. 1, p. 119, Dec. 2010, doi: 10.1186/1471-2105-11-119.

[391] R. D. Finn *et al.*, "InterPro in 2017—beyond protein family and domain annotations," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D190–D199, Jan. 2017, doi: 10.1093/nar/gkw1107.

[392] "SPARQL Endpoint interface to Python." https://rdflib.github.io/sparqlwrapper/ (accessed Aug. 08, 2018).

[393] Duncan Temple Lang and the CRAN team, "CRAN - Package RCurl." Accessed: Nov.

01, 2018. [Online]. Available: https://cran.r-project.org/web/packages/RCurl/index.html

[394] L. Snipen and K. H. Liland, "micropan: an R-package for microbial pan-genomics," *BMC Bioinformatics*, vol. 16, no. 1, p. 79, Dec. 2015, doi: 10.1186/s12859-015-0517-0.

[395] G. Fang, E. Rocha, and A. Danchin, "How Essential Are Nonessential Genes?," *Mol. Biol. Evol.*, vol. 22, no. 11, pp. 2147–2156, Nov. 2005, doi: 10.1093/molbev/msi211.

[396] M. T. Christiansen, R. S. Kaas, R. R. Chaudhuri, M. A. Holmes, H. Hasman, and F. M. Aarestrup, "Genome-Wide High-Throughput Screening to Investigate Essential Genes Involved in Methicillin-Resistant Staphylococcus aureus Sequence Type 398 Survival," *PLoS One*, vol. 9, no. 2, p. e89018, Feb. 2014, doi: 10.1371/journal.pone.0089018.

[397] P. D. Fey *et al.*, "A Genetic Resource for Rapid and Comprehensive Phenotype Screening of Nonessential Staphylococcus aureus Genes," *MBio*, vol. 4, no. 1, pp. 1–8, Mar. 2013, doi: 10.1128/mBio.00537-12.

[398] Y. Le Breton *et al.*, "Essential Genes in the Core Genome of the Human Pathogen Streptococcus pyogenes," *Sci. Rep.*, vol. 5, no. 1, p. 9838, Sep. 2015, doi: 10.1038/srep09838.

[399] J. Levering *et al.*, "Genome-scale reconstruction of the Streptococcus pyogenes M49 metabolic network reveals growth requirements and indicates potential drug targets," *J. Biotechnol.*, vol. 232, pp. 25–37, Aug. 2016, doi: 10.1016/j.jbiotec.2016.01.035.

[400] L. L. C. Gurobi Optimization, "Gurobi Optimizer Reference Manual." 2018. [Online]. Available: http://www.gurobi.com

[401] A. Ebrahim, J. A. Lerman, B. O. Palsson, and D. R. Hyduke, "COBRApy: COnstraints-Based Reconstruction and Analysis for Python," *BMC Syst. Biol.*, vol. 7, no. 1, p. 74, Dec. 2013, doi: 10.1186/1752-0509-7-74.

[402] M. Gostev *et al.*, "The BioSample Database (BioSD) at the European Bioinformatics Institute," *Nucleic Acids Res.*, vol. 40, no. D1, pp. D64–D70, Jan. 2012, doi: 10.1093/nar/gkr937.

[403] B. Zhang, J. Zhang, and L. Sun, "Streptococcus iniae SF1: Complete Genome Sequence, Proteomic Profile, and Immunoprotective Antigens," *PLoS One*, vol. 9, no. 3, p. e91324, Mar. 2014, doi: 10.1371/journal.pone.0091324.

[404] J. W. Pridgeon, D. Zhang, and L. Zhang, "Complete Genome Sequence of the Attenuated Novobiocin-Resistant Streptococcus iniae Vaccine Strain ISNO," *Genome Announc.*, vol. 2, no. 3, pp. 2007–2008, Jun. 2014, doi: 10.1128/genomeA.00510-14.

[405] J.-R. Sun, J.-C. Yan, C.-Y. Yeh, S.-Y. Lee, and J.-J. Lu, "Invasive infection with Streptococcus iniae in Taiwan," *J. Med. Microbiol.*, vol. 56, no. 9, pp. 1246–1249, Sep. 2007, doi: 10.1099/jmm.0.47180-0.

[406] S. Rajoo *et al.*, "Complete Genome Sequence of Streptococcus iniae YSFST01-82, Isolated from Olive Flounder in Jeju, South Korea," *Genome Announc.*, vol. 3, no. 2, pp. 10–11, Apr. 2015, doi: 10.1128/genomeA.00319-15.

[407] M. T. G. Holden *et al.*, "Genomic Evidence for the Evolution of Streptococcus equi: Host Restriction, Increased Virulence, and Genetic Exchange with Human Pathogens," *PLoS Pathog.*, vol. 5, no. 3, p. e1000346, Mar. 2009, doi: 10.1371/journal.ppat.1000346.

[408] S. Pelkonen *et al.*, "Transmission of Streptococcus equi Subspecies zooepidemicus Infection from Horses to Humans," *Emerg. Infect. Dis.*, vol. 19, no. 7, pp. 1041–1048,

Jul. 2013, doi: 10.3201/eid1907.121365.

[409] Z. Ma *et al.*, "Complete Genome Sequence of Streptococcus equi subsp. zooepidemicus Strain ATCC 35246," *J. Bacteriol.*, vol. 193, no. 19, pp. 5583–5584, Oct. 2011, doi: 10.1128/JB.05700-11.

[410] Y. Zhang *et al.*, "Effect of the glycosyltransferases on the capsular polysaccharide synthesis of Streptococcus suis serotype 2," *Microbiol. Res.*, vol. 185, pp. 45–54, Apr. 2016, doi: 10.1016/j.micres.2016.02.002.

[411] Y. Zhang *et al.*, "SssP1, a Streptococcus suis Fimbria-Like Protein Transported by the SecY2/A2 System, Contributes to Bacterial Virulence," *Appl. Environ. Microbiol.*, vol. 84, no. 18, pp. 1–17, Sep. 2018, doi: 10.1128/AEM.01385-18.

[412] C. Chen *et al.*, "A Glimpse of Streptococcal Toxic Shock Syndrome from Comparative Genomics of S. suis 2 Chinese Isolates," *PLoS One*, vol. 2, no. 3, p. e315, Mar. 2007, doi: 10.1371/journal.pone.0000315.

[413] T. Barrett *et al.*, "BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata," *Nucleic Acids Res.*, vol. 40, no. D1, pp. D57–D63, Jan. 2012, doi: 10.1093/nar/gkr1163.

[414] Z. Pan *et al.*, "Novel Variant Serotype of Streptococcus suis Isolated from Piglets with Meningitis," *Appl. Environ. Microbiol.*, vol. 81, no. 3, pp. 976–985, Feb. 2015, doi: 10.1128/AEM.02962-14.

[415] A. Zhang *et al.*, "Comparative genomic analysis of Streptococcus suis reveals significant genomic diversity among different serotypes," *BMC Genomics*, vol. 12, no. 1, p. 523, Dec. 2011, doi: 10.1186/1471-2164-12-523.

[416] S. Chatellier *et al.*, "Phylogenetic diversity of Streptococcus suis strains of various serotypes as revealed by 16S rRNA gene sequence comparison," *Int. J. Syst. Bacteriol.*, vol. 48, no. 2, pp. 581–589, Apr. 1998, doi: 10.1099/00207713-48-2-581.

[417] X. Yao *et al.*, "Isolation and characterization of a native avirulent strain of Streptococcus suis serotype 2: a perspective for vaccine development," *Sci. Rep.*, vol. 5, no. 1, p. 9835, Sep. 2015, doi: 10.1038/srep09835.

[418] B. Boyle, K. Vaillancourt, L. Bonifait, S. J. Charette, M. Gottschalk, and D. Grenier, "Genome Sequence of the Swine Pathogen Streptococcus suis Serotype 2 Strain S735," *J. Bacteriol.*, vol. 194, no. 22, pp. 6343–6344, Nov. 2012, doi: 10.1128/JB.01559-12.

[419] P. Hu *et al.*, "Complete Genome Sequence of Streptococcus suis Serotype 14 Strain JS14," *J. Bacteriol.*, vol. 193, no. 9, pp. 2375–2376, May 2011, doi: 10.1128/JB.00083-11.

[420] H. Zheng *et al.*, "Genomic comparisons of Streptococcus suis serotype 9 strains recovered from diseased pigs in Spain and Canada," *Vet. Res.*, vol. 49, no. 1, p. 1, Dec. 2018, doi: 10.1186/s13567-017-0498-2.

[421] K. Wang, J. Chen, H. Yao, and C. Lu, "Whole-Genome Sequence of Streptococcus suis Serotype 4 Reference Strain 6407," *Genome Announc.*, vol. 2, no. 4, pp. 9–10, Aug. 2014, doi: 10.1128/genomeA.00770-14.

[422] S. P. Szafrański *et al.*, "Quorum sensing of Streptococcus mutans is activated by Aggregatibacter actinomycetemcomitans and by the periodontal microbiome," *BMC Genomics*, vol. 18, no. 1, p. 238, Dec. 2017, doi: 10.1186/s12864-017-3618-5.

[423] L. C. Cook, B. LaSarre, and M. J. Federle, "Interspecies Communication among Commensal and Pathogenic Streptococci," *MBio*, vol. 4, no. 4, pp. 1–11, Aug. 2013, doi: 10.1128/mBio.00382-13.

[424] S. Brouwer *et al.*, "Endopeptidase PepO Regulates the SpeB Cysteine Protease and Is Essential for the Virulence of Invasive M1T1 Streptococcus pyogenes," *J. Bacteriol.*, vol. 200, no. 8, p. JB.00654-17, Apr. 2018, doi: 10.1128/JB.00654-17.

[425] G. Y. C. Cheung and M. Otto, "Understanding the significance of Staphylococcus epidermidis bacteremia in babies and children," *Curr. Opin. Infect. Dis.*, vol. 23, no. 3, pp. 208–216, Jun. 2010, doi: 10.1097/QCO.0b013e328337fecb.

[426] P. Herman-Bausier, C. Labate, A. M. Towell, S. Derclaye, J. A. Geoghegan, and Y. F. Dufrêne, "Staphylococcus aureus clumping factor A is a force-sensitive molecular switch that activates bacterial adhesion," *Proc. Natl. Acad. Sci.*, vol. 115, no. 21, pp. 5564–5569, May 2018, doi: 10.1073/pnas.1718104115.

[427] M. Carlson, "GO.db: A set of annotation maps describing the entire Gene Ontology." 2017.

[428] T. Galili, "dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering," *Bioinformatics*, vol. 31, no. 22, pp. 3718–3720, Nov. 2015, doi: 10.1093/bioinformatics/btv428.

[429] T. Galili, A. O'Callaghan, J. Sidi, and C. Sievert, "heatmaply: an R package for creating interactive cluster heatmaps for online publishing," *Bioinformatics*, vol. 34, no. 9, pp. 1600–1602, May 2018, doi: 10.1093/bioinformatics/btx657.

[430] R Core Team, "R: A Language and Environment for Statistical Computing." Vienna, Austria, 2017. [Online]. Available: https://www.r-project.org/

[431] C. R. García-Alonso, L. M. Pérez-Naranjo, and J. C. Fernández-Caballero, "Multiobjective evolutionary algorithms to identify highly autocorrelated areas: the case of spatial distribution in financially compromised farms," *Ann. Oper. Res.*, vol. 219, no. 1, pp. 187–202, Aug. 2014, doi: 10.1007/s10479-011-0841-3.

[432] C. Ginestet, "ggplot2: Elegant Graphics for Data Analysis," *J. R. Stat. Soc. Ser. A (Statistics Soc.*, vol. 174, no. 1, pp. 245–246, Jan. 2011, doi: 10.1111/j.1467-985X.2010.00676_9.x.

[433] N. A. Moran, "Microbial Minimalism," *Cell*, vol. 108, no. 5, pp. 583–586, Mar. 2002, doi: 10.1016/S0092-8674(02)00665-7.

[434] J. P. McCutcheon and N. A. Moran, "Extreme genome reduction in symbiotic bacteria," *Nat. Rev. Microbiol.*, vol. 10, no. 1, pp. 13–26, 2012, doi: 10.1038/nrmicro2670.

[435] J. I. Glass *et al.*, "Essential genes of a minimal bacterium," *Proc. Natl. Acad. Sci.*, vol. 103, no. 2, pp. 425–430, Jan. 2006, doi: 10.1073/pnas.0510013103.

[436] M. Breuer *et al.*, "Essential metabolism for a minimal cell," *Elife*, vol. 8, pp. 1–75, Jan. 2019, doi: 10.7554/eLife.36842.

[437] E. Yus *et al.*, "Impact of Genome Reduction on Bacterial Metabolism and Its Regulation," *Science (80-.).*, vol. 326, no. 5957, pp. 1263–1268, Nov. 2009, doi: 10.1126/science.1177263.

[438] E. Gaspari, "Model-driven design of Mycoplasma as a vaccine chassis," Wageningen University, 2021. doi: 10.18174/539593.

[439] K. B. Waites, L. Xiao, Y. Liu, M. F. Balish, and T. P. Atkinson, "Mycoplasma pneumoniae from the Respiratory Tract and Beyond," *Clin. Microbiol. Rev.*, vol. 30, no. 3, pp. 747–809, Jul. 2017, doi: 10.1128/CMR.00114-16.

[440] Z. Jiang, S. Li, C. Zhu, R. Zhou, and P. H. M. Leung, "Mycoplasma pneumoniae

Infections: Pathogenesis and Vaccine Development," *Pathogens*, vol. 10, no. 2, p. 119, Jan. 2021, doi: 10.3390/pathogens10020119.

[441] C. Hames, S. Halbedel, M. Hoppert, J. Frey, and J. Stülke, "Glycerol Metabolism Is Important for Cytotoxicity of Mycoplasma pneumoniae," *J. Bacteriol.*, vol. 191, no. 3, pp. 747–753, Feb. 2009, doi: 10.1128/JB.01103-08.

[442] M. Courtot *et al.*, "BioSamples database: an updated sample metadata hub," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D1172–D1178, Jan. 2019, doi: 10.1093/nar/gky1061.

[443] W. L. Simmons, J. M. Daubenspeck, J. D. Osborne, M. F. Balish, K. B. Waites, and K. Dybvig, "Type 1 and type 2 strains of Mycoplasma pneumoniae form different biofilms," *Microbiology*, vol. 159, no. Pt_4, pp. 737–747, Apr. 2013, doi: 10.1099/mic.0.064782-0.

[444] M. Feng, A. C. Burgess, R. R. Cuellar, N. R. Schwab, and M. F. Balish, "Modelling persistent Mycoplasma pneumoniae biofilm infections in a submerged BEAS-2B bronchial epithelial tissue culture model," *J. Med. Microbiol.*, vol. 70, no. 1, Jan. 2021, doi: 10.1099/jmm.0.001266.

[445] J. Galante, A. Ho, S. Tingey, and B. Charalambous, "Quorum Sensing and Biofilms in the Pathogen, Streptococcus pneumoniae," *Curr. Pharm. Des.*, vol. 21, no. 1, pp. 25–30, Nov. 2014, doi: 10.2174/1381612820666140905113336.

[446] L. Chen *et al.*, "Investigation of LuxS-mediated quorum sensing in Klebsiella pneumoniae," *J. Med. Microbiol.*, vol. 69, no. 3, pp. 402–413, Mar. 2020, doi: 10.1099/jmm.0.001148.

[447] D. Balestrino, J. A. J. Haagensen, C. Rich, and C. Forestier, "Characterization of Type 2 Quorum Sensing in Klebsiella pneumoniae and Relationship with Biofilm Formation," *J. Bacteriol.*, vol. 187, no. 8, pp. 2870–2880, Apr. 2005, doi: 10.1128/JB.187.8.2870-2880.2005.

[448] S. Samant *et al.*, "Nucleotide Biosynthesis Is Critical for Growth of Bacteria in Human Blood," *PLoS Pathog.*, vol. 4, no. 2, p. e37, Feb. 2008, doi: 10.1371/journal.ppat.0040037.

[449] J. Barchiesi, M. E. Castelli, G. Di Venanzio, M. I. Colombo, and E. García Véscovi, "The PhoP/PhoQ System and Its Role in Serratia marcescens Pathogenesis," *J. Bacteriol.*, vol. 194, no. 11, pp. 2949–2961, Jun. 2012, doi: 10.1128/JB.06820-11.

[450] S. Chakravarty, C. N. Melton, A. Bailin, T. L. Yahr, and G. G. Anderson, "Pseudomonas aeruginosa Magnesium Transporter MgtE Inhibits Type III Secretion System Gene Expression by Stimulating rsmYZ Transcription," *J. Bacteriol.*, vol. 199, no. 23, pp. 1–11, Dec. 2017, doi: 10.1128/JB.00268-17.

[451] S. Dhandayuthapani, M. W. Blaylock, C. M. Bebear, W. G. Rasmussen, and J. B. Baseman, "Peptide Methionine Sulfoxide Reductase (MsrA) Is a Virulence Determinant in Mycoplasma genitalium," *J. Bacteriol.*, vol. 183, no. 19, pp. 5645–5650, Oct. 2001, doi: 10.1128/JB.183.19.5645-5650.2001.

[452] K. H. Sippel *et al.*, "Insights into Mycoplasma genitalium metabolism revealed by the structure of MG289, an extracytoplasmic thiamine binding lipoprotein," *Proteins Struct. Funct. Bioinforma.*, vol. 79, no. 2, pp. 528–536, Feb. 2011, doi: 10.1002/prot.22900.

[453] J. C. Kaiser and D. E. Heinrichs, "Branching Out: Alterations in Bacterial Physiology and Virulence Due to Branched-Chain Amino Acid Deprivation," *MBio*, vol. 9, no. 5, Nov. 2018, doi: 10.1128/mBio.01188-18.

[454] G. Gesbert *et al.*, "Importance of Branched-Chain Amino Acid Utilization in Francisella Intracellular Adaptation," *Infect. Immun.*, vol. 83, no. 1, pp. 173–183, Jan. 2015, doi: 10.1128/IAI.02579-14.

[455] M. Holeček, "Branched-chain amino acids in health and disease: metabolism, alterations in blood plasma, and as supplements," *Nutr. Metab. (Lond).*, vol. 15, no. 1, p. 33, Dec. 2018, doi: 10.1186/s12986-018-0271-1.

[456] J. Timenetsky, L. M. Santos, M. Buzinhani, and E. Mettifogo, "Detection of multiple mycoplasma infection in cell cultures by PCR," *Brazilian J. Med. Biol. Res.*, vol. 39, no. 7, pp. 907–914, Jul. 2006, doi: 10.1590/S0100-879X2006000700009.

[457] "Mycoplasma pneumoniae Surveillance and Reporting | CDC," 2020. https://www.cdc.gov/pneumonia/atypical/mycoplasma/surv-reporting.html (accessed May 21, 2021).

[458] F. Y. Khan and M. A.yassin, "Mycoplasma pneumoniae associated with severe autoimmune hemolytic anemia: case report and literature review," *Brazilian J. Infect. Dis.*, vol. 13, no. 1, pp. 77–79, Feb. 2009, doi: 10.1590/S1413-86702009000100018.

[459] M. Güell *et al.*, "Transcriptome Complexity in a Genome-Reduced Bacterium," *Science (80-.).*, vol. 326, no. 5957, pp. 1268–1271, Nov. 2009, doi: 10.1126/science.1176951.

[460] T. Maier *et al.*, "Quantification of mRNA and protein and integration with protein turnover in a bacterium," *Mol. Syst. Biol.*, vol. 7, no. 1, p. 511, Jan. 2011, doi: 10.1038/msb.2011.38.

[461] T. Maier *et al.*, "Large-scale metabolome analysis and quantitative integration with genomics and proteomics data in Mycoplasma pneumoniae," *Mol. Biosyst.*, vol. 9, no. 7, p. 1743, 2013, doi: 10.1039/c3mb70113a.

[462] I. Junier, E. B. Unal, E. Yus, V. Lloréns-Rico, and L. Serrano, "Insights into the Mechanisms of Basal Coordination of Transcription Using a Genome-Reduced Bacterium," *Cell Syst.*, vol. 2, no. 6, pp. 391–401, Jun. 2016, doi: 10.1016/j.cels.2016.04.015.

[463] E. Yus *et al.*, "Determination of the Gene Regulatory Network of a Genome-Reduced Bacterium Highlights Alternative Regulation Independent of Transcription Factors," *Cell Syst.*, vol. 9, no. 2, pp. 143-158.e13, Aug. 2019, doi: 10.1016/j.cels.2019.07.001.

[464] H. Kurata and Y. Sugimoto, "Improved kinetic model of Escherichia coli central carbon metabolism in batch and continuous cultures," *J. Biosci. Bioeng.*, vol. 125, no. 2, pp. 251–257, Feb. 2018, doi: 10.1016/j.jbiosc.2017.09.005.

[465] D. Machado, M. J. Herrgård, and I. Rocha, "Modeling the Contribution of Allosteric Regulation for Flux Control in the Central Carbon Metabolism of E. coli," *Front. Bioeng. Biotechnol.*, vol. 3, no. October, pp. 1–11, Oct. 2015, doi: 10.3389/fbioe.2015.00154.

[466] G. M. Oddone, D. a. Mills, and D. E. Block, "A dynamic, genome-scale flux model of Lactococcus lactis to increase specific recombinant protein expression," *Metab. Eng.*, vol. 11, no. 6, pp. 367–381, Nov. 2009, doi: 10.1016/j.ymben.2009.07.007.

[467] P. Millard, K. Smallbone, and P. Mendes, "Metabolic regulation is sufficient for global and robust coordination of glucose uptake, catabolism, energy production and growth in Escherichia coli," *PLOS Comput. Biol.*, vol. 13, no. 2, p. e1005396, Feb. 2017, doi: 10.1371/journal.pcbi.1005396.

[468] K. Smallbone *et al.*, "A model of yeast glycolysis based on a consistent kinetic characterisation of all its enzymes," *FEBS Lett.*, vol. 587, no. 17, pp. 2832–2841, Sep.

2013, doi: 10.1016/j.febslet.2013.06.043.

[469] E. Murabito *et al.*, "Monte-Carlo Modeling of the Central Carbon Metabolism of Lactococcus lactis: Insights into Metabolic Regulation," *PLoS One*, vol. 9, no. 9, p. e106453, Sep. 2014, doi: 10.1371/journal.pone.0106453.

[470] T. Kamminga, S. Slagman, J. J. E. Bijlsma, V. A. P. Martins dos Santos, M. Suarez-Diez, and P. J. Schaap, "Metabolic modeling of energy balances in Mycoplasma hyopneumoniae shows that pyruvate addition increases growth rate," *Biotechnol. Bioeng.*, vol. 114, no. 10, pp. 2339–2347, Oct. 2017, doi: 10.1002/bit.26347.

[471] D. Kesten, U. Kummer, S. Sahle, and K. Hübner, "A new model for the aerobic metabolism of yeast allows the detailed analysis of the metabolic regulation during glucose pulse," *Biophys. Chem.*, vol. 206, pp. 40–57, Nov. 2015, doi: 10.1016/j.bpc.2015.06.010.

[472] K. Kochanowski, L. Gerosa, S. F. Brunner, D. Christodoulou, Y. V Nikolaev, and U. Sauer, "Few regulatory metabolites coordinate expression of central metabolic genes in Escherichia coli," *Mol. Syst. Biol.*, vol. 13, no. 1, p. 903, Jan. 2017, doi: 10.15252/msb.20167402.

[473] S. Halbedel and J. Stülke, "Probing In Vivo Promoter Activities in Mycoplasma pneumoniae : a System for Generation of Single-Copy Reporter Constructs," *Appl. Environ. Microbiol.*, vol. 72, no. 2, pp. 1696–1699, Feb. 2006, doi: 10.1128/AEM.72.2.1696-1699.2006.

[474] S. Halbedel, J. Busse, S. R. Schmidl, and J. Stülke, "Regulatory Protein Phosphorylation in Mycoplasma pneumoniae," *J. Biol. Chem.*, vol. 281, no. 36, pp. 26253–26259, Sep. 2006, doi: 10.1074/jbc.M605010200.

[475] T. Fuhrer, D. Heer, B. Begemann, and N. Zamboni, "High-Throughput, Accurate Mass Metabolome Profiling of Cellular Extracts by Flow Injection–Time-of-Flight Mass Spectrometry," *Anal. Chem.*, vol. 83, no. 18, pp. 7074–7080, Sep. 2011, doi: 10.1021/ac201267k.

[476] J. D. Storey and R. Tibshirani, "Statistical significance for genomewide studies," *Proc. Natl. Acad. Sci.*, vol. 100, no. 16, pp. 9440–9445, Aug. 2003, doi: 10.1073/pnas.1530509100.

[477] J. M. Buescher, S. Moco, U. Sauer, and N. Zamboni, "Ultrahigh Performance Liquid Chromatography–Tandem Mass Spectrometry Method for Fast and Robust Quantification of Anionic and Aromatic Metabolites," *Anal. Chem.*, vol. 82, no. 11, pp. 4403–4412, Jun. 2010, doi: 10.1021/ac100101d.

[478] K. Wolstencroft *et al.*, "The SEEK," in *Methods in Enzymology*, vol. 500, 2011, pp. 629–655. doi: 10.1016/B978-0-12-385118-5.00029-3.

[479] T. Lubitz, J. Hahn, F. T. Bergmann, E. Noor, E. Klipp, and W. Liebermeister, "SBtab: a flexible table format for data exchange in systems biology," *Bioinformatics*, vol. 32, no. 16, pp. 2559–2561, Aug. 2016, doi: 10.1093/bioinformatics/btw179.

[480] N. Juty, N. Le Novere, and C. Laibe, "Identifiers.org and MIRIAM Registry: community resources to provide persistent identification," *Nucleic Acids Res.*, vol. 40, no. D1, pp. D580–D586, Jan. 2012, doi: 10.1093/nar/gkr1097.

[481] M. Courtot *et al.*, "Controlled vocabularies and semantics in systems biology," *Mol. Syst. Biol.*, vol. 7, no. 1, p. 543, Jan. 2011, doi: 10.1038/msb.2011.77.

[482] A. Jauhiainen, B. Madhu, M. Narita, M. Narita, J. Griffiths, and S. Tavaré, "Normalization of metabolomics data with applications to correlation maps,"

Bioinformatics, vol. 30, no. 15, pp. 2155–2161, Aug. 2014, doi: 10.1093/bioinformatics/btu175.

[483] Y. Wu and L. Li, "Sample normalization methods in quantitative metabolomics," *J. Chromatogr. A*, vol. 1430, pp. 80–95, Jan. 2016, doi: 10.1016/j.chroma.2015.12.007.

[484] R. S. Costa, A. Hartmann, P. Gaspar, A. R. Neves, and S. Vinga, "An extended dynamic model of Lactococcus lactis metabolism for mannitol and 2,3-butanediol production," *Mol. Biosyst.*, vol. 10, no. 3, p. 628, 2014, doi: 10.1039/c3mb70265k.

[485] S. Hoops *et al.*, "COPASI--a COmplex PAthway SImulator," *Bioinformatics*, vol. 22, no. 24, pp. 3067–3074, Dec. 2006, doi: 10.1093/bioinformatics/btl485.

[486] A. Flamholz, E. Noor, A. Bar-Even, and R. Milo, "eQuilibrator--the biochemical thermodynamics calculator," *Nucleic Acids Res.*, vol. 40, no. D1, pp. D770–D775, Jan. 2012, doi: 10.1093/nar/gkr874.

[487] C. Chen, J. Twycross, and J. M. Garibaldi, "A new accuracy measure based on bounded relative error for time series forecasting," *PLoS One*, vol. 12, no. 3, p. e0174202, Mar. 2017, doi: 10.1371/journal.pone.0174202.

[488] R. L. Iman, J. C. Helton, and J. E. Campbell, "An Approach to Sensitivity Analysis of Computer Models: Part I—Introduction, Input Variable Selection and Preliminary Variable Assessment," *J. Qual. Technol.*, vol. 13, no. 3, pp. 174–183, Jul. 1981, doi: 10.1080/00224065.1981.11978748.

[489] R. L. Iman, J. C. Helton, and J. E. Campbell, "An Approach to Sensitivity Analysis of Computer Models: Part II—Ranking of Input Variables, Response Surface Validation, Distribution Effect and Technique Synopsis," *J. Qual. Technol.*, vol. 13, no. 4, pp. 232–240, Oct. 1981, doi: 10.1080/00224065.1981.11978763.

[490] K. Choi, J. K. Medley, C. Cannistra, and K. Matthias, "Tellurium : A Python Based Modeling and Reproducibility Platform for Systems Biology," *bioRxiv*, 2016, doi: https://doi.org/10.1101/054601.

[491] E. T. Somogyi *et al.*, "libRoadRunner: a high performance SBML simulation and analysis library: Table 1.," *Bioinformatics*, vol. 31, no. 20, pp. 3315–3321, Oct. 2015, doi: 10.1093/bioinformatics/btv363.

[492] "Merk TTP T300 tissue culture flask specifications," *TPP® tissue culture flasks | Sigma-Aldrich*. www.sigmaaldrich.com/catalog/product/sigma/z707562?lang=en&region=NL (accessed Jul. 12, 2018).

[493] C. R. Wilke and P. Chang, "Correlation of diffusion coefficients in dilute solutions," *AIChE J.*, vol. 1, no. 2, pp. 264–270, Jun. 1955, doi: 10.1002/aic.690010222.

[494] H. J. V. Tyrrell, "The origin and present status of Fick's diffusion law," *J. Chem. Educ.*, vol. 41, no. 7, p. 397, Jul. 1964, doi: 10.1021/ed041p397.

[495] T. Tiernan, C. Chang, and C. C. Cheng, "Formation and Reactions of Negative Ions Relevant to Chemical Ionization Mass Spectrometry . I . Cl Mass Spectra of Organic Compounds Produced by F- Reactions," *Environ. Health Perspect.*, vol. 36, pp. 47–62, 1980, doi: https://doi.org/10.1371/journal.pbio.2001414.

[496] M. Yang and K. H. Vousden, "Serine and one-carbon metabolism in cancer," *Nat. Rev. Cancer*, vol. 16, no. 10, pp. 650–662, Oct. 2016, doi: 10.1038/nrc.2016.81.

[497] S. R. Schmidl *et al.*, "A Trigger Enzyme in Mycoplasma pneumoniae: Impact of the Glycerophosphodiesterase GlpQ on Virulence and Gene Expression," *PLoS Pathog.*, vol. 7, no. 9, p. e1002263, Sep. 2011, doi: 10.1371/journal.ppat.1002263.

[498] J. Joseph and J. Loscalzo, "Methoxistasis: Integrating the Roles of Homocysteine and Folic Acid in Cardiovascular Pathobiology," *Nutrients*, vol. 5, no. 8, pp. 3235–3256, Aug. 2013, doi: 10.3390/nu5083235.

[499] S. Großhennig, S. R. Schmidl, G. Schmeisky, J. Busse, and J. Stülke, "Implication of Glycerol and Phospholipid Transporters in Mycoplasma pneumoniae Growth and Virulence," *Infect. Immun.*, vol. 81, no. 3, pp. 896–904, Mar. 2013, doi: 10.1128/IAI.01212-12.

[500] N. Le Novère *et al.*, "The Systems Biology Graphical Notation," *Nat. Biotechnol.*, vol. 27, no. 8, pp. 735–741, Aug. 2009, doi: 10.1038/nbt.1558.

[501] V. Chelliah *et al.*, "BioModels: Ten-year anniversary," *Nucleic Acids Res.*, vol. 43, no. D1, pp. D542–D548, 2015, doi: 10.1093/nar/gku1181.

[502] A. Ghorbaniaghdam, O. Henry, and M. Jolicoeur, "An in-silico study of the regulation of CHO cells glycolysis," *J. Theor. Biol.*, vol. 357, pp. 112–122, Sep. 2014, doi: 10.1016/j.jtbi.2014.04.035.

[503] E. Gehrmann, C. Gläßer, Y. Jin, B. Sendhoff, B. Drossel, and K. Hamacher, "Robustness of glycolysis in yeast to internal and external noise," *Phys. Rev. E*, vol. 84, no. 2, p. 021913, Aug. 2011, doi: 10.1103/PhysRevE.84.021913.

[504] A. R. Neves *et al.*, "Is the Glycolytic Flux in Lactococcus lactisPrimarily Controlled by the Redox Charge?," *J. Biol. Chem.*, vol. 277, no. 31, pp. 28088–28098, Aug. 2002, doi: 10.1074/jbc.M202573200.

[505] A. NEVES, W. POOL, J. KOK, O. KUIPERS, and H. SANTOS, "Overview on sugar metabolism and its control in – The input from in vivo NMR," *FEMS Microbiol. Rev.*, vol. 29, no. 3, pp. 531–554, Aug. 2005, doi: 10.1016/j.femsre.2005.04.005.

[506] N. J. DeVito and B. Goldacre, "Catalogue of bias: publication bias," *BMJ Evidence-Based Med.*, vol. 24, no. 2, pp. 53–54, Apr. 2019, doi: 10.1136/bmjebm-2018-111107.

[507] A. Mlinarić, M. Horvat, and V. Šupak Smolčić, "Dealing with the positive publication bias: Why you should really publish your negative results," *Biochem. Medica*, vol. 27, no. 3, pp. 1–6, Oct. 2017, doi: 10.11613/BM.2017.030201.

[508] R. Bailo, A. Bhatt, and J. A. Aínsa, "Lipid transport in Mycobacterium tuberculosis and its implications in virulence and drug development," *Biochem. Pharmacol.*, vol. 96, no. 3, pp. 159–167, Aug. 2015, doi: 10.1016/j.bcp.2015.05.001.

[509] P. Singh, N. R. Rameshwaram, S. Ghosh, and S. Mukhopadhyay, "Cell envelope lipids in the pathophysiology of Mycobacterium tuberculosis," *Future Microbiol.*, vol. 13, no. 6, pp. 689–710, May 2018, doi: 10.2217/fmb-2017-0135.

[510] J. Goodman, R. Wait, and T. Battle, "Polar Lipid Profiling of Mycoplasma Pneumoniae-Infected Human Lung Epithelial Cells," in *New Developments and New Applications in Animal Cell Technology*, Dordrecht: Kluwer Academic Publishers, 1998, pp. 713–715. doi: 10.1007/0-306-46860-3_130.

[511] S. Razin, S. Kutner, H. Efrati, and S. Rottem, "Phospholipid and cholesterol uptake by mycoplasma cells and membranes," *Biochim. Biophys. Acta - Biomembr.*, vol. 598, no. 3, pp. 628–640, Jun. 1980, doi: 10.1016/0005-2736(80)90042-5.

[512] J. D. Pollack, M. V Williams, and R. N. McElhaney, "The Comparative Metabolism of the Mollicutes (Mycoplasmas): The Utility for Taxonomic Classification and the Relationship of Putative Gene Annotation and Phylogeny to Enzymatic Function in the Smallest Free-Living Cells," *Crit. Rev. Microbiol.*, vol. 23, no. 4, pp. 269–354, Jan. 1997, doi: 10.3109/10408419709115140.

[513] D. S. Jordan, J. M. Daubenspeck, A. H. Laube, M. B. Renfrow, and K. Dybvig, "O -linked protein glycosylation in mycoplasma," *Mol. Microbiol.*, vol. 90, no. 5, pp. 1046–1053, Dec. 2013, doi: 10.1111/mmi.12415.

[514] B. L. Beckman and G. E. Kenny, "Immunochemical Analysis of Serologically Active Lipids of Mycoplasma pneumoniae," *J. Bacteriol.*, vol. 96, no. 4, pp. 1171–1180, Oct. 1968, doi: 10.1128/jb.96.4.1171-1180.1968.

[515] A. K. Pandey and C. M. Sassetti, "Mycobacterial persistence requires the utilization of host cholesterol," *Proc. Natl. Acad. Sci.*, vol. 105, no. 11, pp. 4376–4380, Mar. 2008, doi: 10.1073/pnas.0711159105.

[516] K. Bhargava, G. Nath, G. K. Aseri, and N. Jain, "Potential of Bacteriophage Therapy: A Double Edge Sword to Combat COVID-19 and Associated Pulmonary Bacterial Infections," *Indian J. Pharm. Sci.*, vol. 83, no. 6, pp. 1081–1093, 2021, doi: 10.36468/pharmaceutical-sciences.864.

[517] M. I. Bukrinsky, N. Mukhamedova, and D. Sviridov, "Lipid rafts and pathogens: the art of deception and exploitation," *J. Lipid Res.*, vol. 61, no. 5, pp. 601–610, May 2020, doi: 10.1194/jlr.TR119000391.

[518] S. T. Abedon, "Phage-Antibiotic Combination Treatments: Antagonistic Impacts of Antibiotics on the Pharmacodynamics of Phage Therapy?," *Antibiotics*, vol. 8, no. 4, p. 182, Oct. 2019, doi: 10.3390/antibiotics8040182.

[519] D. G. Russell, P. Cardona, M. Kim, S. Allain, and F. Altare, "Foamy macrophages and the progression of the human tuberculosis granuloma," *Nat. Immunol.*, vol. 10, no. 9, pp. 943–948, Sep. 2009, doi: 10.1038/ni.1781.

[520] F. Peyrusson, T. K. Nguyen, T. Najdovski, and F. Van Bambeke, "Host Cell Oxidative Stress Induces Dormant Staphylococcus aureus Persisters," *Microbiol. Spectr.*, vol. 10, no. 1, pp. 1–15, Feb. 2022, doi: 10.1128/spectrum.02313-21.

[521] B. Pascoe *et al.*, "Dormant Cells of Staphylococcus aureus Are Resuscitated by Spent Culture Supernatant," *PLoS One*, vol. 9, no. 2, p. e85998, Feb. 2014, doi: 10.1371/journal.pone.0085998.

[522] M. Gelber, F. Babushkin, and A. Schattner, "Stubborn Creatures: Dormant Staphylococcus aureus," *Am. J. Med.*, vol. 130, no. 3, pp. e101–e102, Mar. 2017, doi: 10.1016/j.amjmed.2016.10.012.

[523] M. Currás, B. Magariños, A. Toranzo, and J. Romalde, "Dormancy as a survival strategy of the fish pathogen Streptococcus parauberis in the marine environment," *Dis. Aquat. Organ.*, vol. 52, no. 2, pp. 129–136, 2002, doi: 10.3354/dao052129.

[524] M. Fraunholz and B. Sinha, "Intracellular staphylococcus aureus: Live-in and let die," *Front. Cell. Infect. Microbiol.*, vol. 2, no. April, p. 43, 2012, doi: 10.3389/fcimb.2012.00043.

[525] A. Moldovan and M. J. Fraunholz, "In or out: Phagosomal escape of Staphylococcus aureus," *Cell. Microbiol.*, vol. 21, no. 3, p. e12997, Mar. 2019, doi: 10.1111/cmi.12997.

[526] A. Slomiany, V. L. N. Murty, M. Aono, C. E. Snyder, A. Herp, and B. L. Slomiany, "Lipid composition of tracheobronchial secretions from normal individuals and patients with cystic fibrosis," *Biochim. Biophys. Acta - Lipids Lipid Metab.*, vol. 710, no. 1, pp. 106–111, Jan. 1982, doi: 10.1016/0005-2760(82)90196-5.

[527] T. M. Jarry and A. L. Cheung, "Staphylococcus aureus Escapes More Efficiently from the Phagosome of a Cystic Fibrosis Bronchial Epithelial Cell Line than from Its Normal Counterpart," *Infect. Immun.*, vol. 74, no. 5, pp. 2568–2577, May 2006, doi:

10.1128/IAI.74.5.2568-2577.2006.

[528] D. Pajuelo, U. Tak, L. Zhang, O. Danilchanka, A. D. Tischler, and M. Niederweis, "Toxin secretion and trafficking by Mycobacterium tuberculosis," *Nat. Commun.*, vol. 12, no. 1, p. 6592, Dec. 2021, doi: 10.1038/s41467-021-26925-1.

[529] J. Aguilera *et al.*, "Nα-Acetylation of the virulence factor EsxA is required for mycobacterial cytosolic translocation and virulence," *J. Biol. Chem.*, vol. 295, no. 17, pp. 5785–5794, Apr. 2020, doi: 10.1074/jbc.RA119.012497.

[530] S. A. Ragland and J. C. Kagan, "Cytosolic detection of phagosomal bacteria—Mechanisms underlying PAMP exodus from the phagosome into the cytosol," *Mol. Microbiol.*, vol. 116, no. 6, pp. 1420–1432, Dec. 2021, doi: 10.1111/mmi.14841.

[531] H. Koliwer-Brandl *et al.*, "Distinct <scp> *Mycobacterium marinum* </scp> phosphatases determine pathogen vacuole phosphoinositide pattern, phagosome maturation, and escape to the cytosol," *Cell. Microbiol.*, vol. 21, no. 6, p. e13008, Jun. 2019, doi: 10.1111/cmi.13008.

[532] T. R. Lerner, C. J. Queval, A. Fearns, U. Repnik, G. Griffiths, and M. G. Gutierrez, "Phthiocerol dimycocerosates promote access to the cytosol and intracellular burden of Mycobacterium tuberculosis in lymphatic endothelial cells," *BMC Biol.*, vol. 16, no. 1, p. 1, Dec. 2018, doi: 10.1186/s12915-017-0471-6.

[533] E. S. Seilie and J. Bubeck Wardenburg, "Staphylococcus aureus pore-forming toxins: The interface of pathogen and host complexity," *Semin. Cell Dev. Biol.*, vol. 72, no. 10, pp. 101–116, Dec. 2017, doi: 10.1016/j.semcdb.2017.04.003.

[534] K. van Pee, E. Mulvihill, D. J. Müller, and Ö. Yildiz, "Unraveling the Pore-Forming Steps of Pneumolysin from Streptococcus pneumoniae," *Nano Lett.*, vol. 16, no. 12, pp. 7915–7924, Dec. 2016, doi: 10.1021/acs.nanolett.6b04219.

[535] J. E. Alouf and H. Müller-Alouf, "Staphylococcal and streptococcal superantigens: molecular, biological and clinical aspects," *Int. J. Med. Microbiol.*, vol. 292, no. 7–8, pp. 429–440, 2003, doi: 10.1078/1438-4221-00232.

[536] B. L. Spencer, U. Tak, J. C. Mendonça, P. E. Nagao, M. Niederweis, and K. S. Doran, "A type VII secretion system in Group B Streptococcus mediates cytotoxicity and virulence," *PLOS Pathog.*, vol. 17, no. 12, p. e1010121, Dec. 2021, doi: 10.1371/journal.ppat.1010121.

[537] I. J. Glomski, M. M. Gedde, A. W. Tsang, J. A. Swanson, and D. A. Portnoy, "The Listeria monocytogenes hemolysin has an acidic pH optimum to compartmentalize activity and prevent damage to infected host cells," *J. Cell Biol.*, vol. 156, no. 6, pp. 1029–1038, Mar. 2002, doi: 10.1083/jcb.200201081.

[538] J. Pinheiro *et al.*, "Listeria monocytogenes encodes a functional ESX-1 secretion system whose expression is detrimental to in vivo infection," *Virulence*, vol. 8, no. 6, pp. 993–1004, Aug. 2017, doi: 10.1080/21505594.2016.1244589.

[539] Y. Shi, M. J. Cromie, F.-F. Hsu, J. Turk, and E. A. Groisman, "PhoP-regulated Salmonella resistance to the antimicrobial peptides magainin 2 and polymyxin B," *Mol. Microbiol.*, vol. 53, no. 1, pp. 229–241, May 2004, doi: 10.1111/j.1365-2958.2004.04107.x.

[540] G. Weiss and P. L. Carver, "Role of divalent metals in infectious disease susceptibility and outcome," *Clin. Microbiol. Infect.*, vol. 24, no. 1, pp. 16–23, Jan. 2018, doi: 10.1016/j.cmi.2017.01.018.

[541] X. Li *et al.*, "A combination therapy of Phages and Antibiotics: Two is better than one,"

Int. J. Biol. Sci., vol. 17, no. 13, pp. 3573–3582, 2021, doi: 10.7150/ijbs.60551.

[542] D. M. Lin, B. Koskella, and H. C. Lin, "Phage therapy: An alternative to antibiotics in the age of multi-drug resistance," *World J. Gastrointest. Pharmacol. Ther.*, vol. 8, no. 3, p. 162, 2017, doi: 10.4292/wjgpt.v8.i3.162.

[543] A. Abdelsattar *et al.*, "How to Train Your Phage: The Recent Efforts in Phage Training," *Biologics*, vol. 1, no. 2, pp. 70–88, Jul. 2021, doi: 10.3390/biologics1020005.

[544] "Tuberculosis Statistics & Facts (2021 Update) - PolicyAdvice." https://policyadvice.net/insurance/insights/tuberculosis-statistics/ (accessed Jun. 20, 2022).

[545] T. J. Basera, J. Ncayiyana, and M. E. Engel, "Prevalence and risk factors of latent tuberculosis infection in Africa: a systematic review and meta-analysis protocol," *BMJ Open*, vol. 7, no. 7, p. e012636, Jul. 2017, doi: 10.1136/bmjopen-2016-012636.

[546] Z. A. King, A. Dräger, A. Ebrahim, N. Sonnenschein, N. E. Lewis, and B. O. Palsson, "Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways," *PLOS Comput. Biol.*, vol. 11, no. 8, p. e1004321, Aug. 2015, doi: 10.1371/journal.pcbi.1004321.

[547] K.-H. Cheung, K. Y. Yip, A. Smith, R. DeKnikker, A. Masiar, and M. Gerstein, "YeastHub: a semantic web use case for integrating data in the life sciences domain," *Bioinformatics*, vol. 21, no. Suppl 1, pp. i85–i96, Jun. 2005, doi: 10.1093/bioinformatics/bti1026.

[548] A. Waagmeester, "First steps towards WikiPathways RDF," *Nat. Preced.*, Aug. 2011, doi: 10.1038/npre.2011.6300.1.

[549] A. Waagmeester *et al.*, "Using the Semantic Web for Rapid Integration of WikiPathways with Other Biological Online Data Resources," *PLOS Comput. Biol.*, vol. 12, no. 6, p. e1004989, Jun. 2016, doi: 10.1371/journal.pcbi.1004989.

[550] K. Forslund and E. L. L. Sonnhammer, "Predicting protein function from domain content," *Bioinformatics*, vol. 24, no. 15, pp. 1681–1687, Aug. 2008, doi: 10.1093/bioinformatics/btn312.

[551] B. Teusink, H. Bachmann, and D. Molenaar, "Systems biology of lactic acid bacteria: a critical review," *Microb. Cell Fact.*, vol. 10, no. Suppl 1, p. S11, 2011, doi: 10.1186/1475-2859-10-S1-S11.

[552] D. Ganguly, "A Fast Partitional Clustering Algorithm based on Nearest Neighbours Heuristics," *Pattern Recognit. Lett.*, vol. 112, pp. 198–204, Sep. 2018, doi: 10.1016/j.patrec.2018.07.017.

[553] N. Murugesan, I. Cho, and C. Tortora, "Benchmarking in Cluster Analysis: A Study on Spectral Clustering, DBSCAN, and K-Means," in *Data Analysis and Rationality in a Complex World*, 2021, pp. 175–185.

[554] B. Worley, S. Halouska, and R. Powers, "Utilities for quantifying separation in PCA/PLS-DA scores plots," *Anal. Biochem.*, vol. 433, no. 2, pp. 102–104, Feb. 2013, doi: 10.1016/j.ab.2012.10.011.

[555] R. Motta, R. Minghim, A. de Andrade Lopes, and M. C. F. Oliveira, "Graph-based measures to assist user assessment of multidimensional projections," *Neurocomputing*, vol. 150, no. PB, pp. 583–598, Feb. 2015, doi: 10.1016/j.neucom.2014.09.063.

[556] S. Z. Alborzi, M.-D. Devignes, and D. W. Ritchie, *Bioinformatics and Biomedical*

Engineering, vol. 10209. Cham: Springer International Publishing, 2017. doi: 10.1007/978-3-319-56154-7.

[557] R. D. Fleischmann *et al.*, "Whole-Genome Random Sequencing and Assembly of Haemophilus influenzae Rd," *Science (80-.).*, vol. 269, no. 5223, pp. 496–512, Jul. 1995, doi: 10.1126/science.7542800.

[558] A. Funahashi, Y. Matsuoka, A. Jouraku, H. Kitano, and N. Kikuchi, "Celldesigner: A Modeling Tool for Biochemical Networks," in *Proceedings of the 2006 Winter Simulation Conference*, Dec. 2006, vol. 1, no. 5, pp. 1707–1712. doi: 10.1109/WSC.2006.322946.

[559] J. H. Harris, "Chapter 18," in *N*, vol. 1696, Fortress Press, 2021, pp. 177–180. doi: 10.2307/j.ctv1khdp5v.21.

[560] C. M. Welsh *et al.*, "PyCoTools: a Python toolbox for COPASI," *Bioinformatics*, vol. 34, no. 21, pp. 3702–3710, Nov. 2018, doi: 10.1093/bioinformatics/bty409.

[561] J. K. Medley *et al.*, "Tellurium notebooks—An environment for reproducible dynamical modeling in systems biology," *PLOS Comput. Biol.*, vol. 14, no. 6, p. e1006220, Jun. 2018, doi: 10.1371/journal.pcbi.1006220.

[562] "European Open Science Cloud," *Nat. Genet.*, vol. 48, no. 8, pp. 821–821, Aug. 2016, doi: 10.1038/ng.3642.

[563] E. Schultes, G. Strawn, and B. Mons, "Ready, set, go fair: Accelerating convergence to an internet of fair data and services," *CEUR Workshop Proc.*, vol. 2277, pp. 19–23, 2018, [Online]. Available: http://ceur-ws.org/Vol-2277/paper07.pdf

[564] M. Glont *et al.*, "BioModels: expanding horizons to include more modelling approaches and formats," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D1248–D1253, Jan. 2018, doi: 10.1093/nar/gkx1023.

[565] N. Pham, R. van Heck, J. van Dam, P. Schaap, E. Saccenti, and M. Suarez-Diez, "Consistency, Inconsistency, and Ambiguity of Metabolite Names in Biochemical Databases Used for Genome-Scale Metabolic Modelling," *Metabolites*, vol. 9, no. 2, p. 28, Feb. 2019, doi: 10.3390/metabo9020028.

[566] S. Moretti, V. D. T. Tran, F. Mehl, M. Ibberson, and M. Pagni, "MetaNetX/MNXref: unified namespace for metabolites and biochemical reactions in the context of metabolic models," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D570–D574, Jan. 2021, doi: 10.1093/nar/gkaa992.

[567] "chemspider." https://www.chemspider.com/

[568] S. Kim *et al.*, "PubChem 2019 update: improved access to chemical data," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D1102–D1109, Jan. 2019, doi: 10.1093/nar/gky1033.

[569] A. Waagmeester *et al.*, "Wikidata as a knowledge graph for the life sciences," *Elife*, vol. 9, pp. 1–15, Mar. 2020, doi: 10.7554/eLife.52614.

[570] J. Hastings *et al.*, "ChEBI in 2016: Improved services and an expanding collection of metabolites," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D1214–D1219, Jan. 2016, doi: 10.1093/nar/gkv1031.

[571] B. Delépine, T. Duigou, P. Carbonell, and J.-L. Faulon, "RetroPath2.0: A retrosynthesis workflow for metabolic engineers," *Metab. Eng.*, vol. 45, no. April 2017, pp. 158–170, Jan. 2018, doi: 10.1016/j.ymben.2017.12.002.

[572] D. Gasevic, "Petri nets on the semantic web guidelines and infrastructure," *Comput. Sci. Inf. Syst.*, vol. 1, no. 2, pp. 127–151, 2004, doi: 10.2298/CSIS0402127G.

[573] W. Marwan, C. Rohr, and M. Heiner, “Petri Nets in Snoopy: A Unifying Framework for the Graphical Display, Computational Modelling, and Simulation of Bacterial Regulatory Networks,” in *Methods in Molecular Biology*, vol. 804, 2012, pp. 409–437. doi: 10.1007/978-1-61779-361-5_21.

[574] C. Chaouiya, “Petri net modelling of biological networks,” *Brief. Bioinform.*, vol. 8, no. 4, pp. 210–219, Mar. 2007, doi: 10.1093/bib/bbm029.

[575] L. Albergante, J. Timmis, L. Beattie, and P. M. Kaye, “A Petri Net Model of Granulomatous Inflammation: Implications for IL-10 Mediated Control of Leishmania donovani Infection,” *PLoS Comput. Biol.*, vol. 9, no. 11, p. e1003334, Nov. 2013, doi: 10.1371/journal.pcbi.1003334.

[576] H. J. Genrich and K. Lautenbach, “System modelling with high-level Petri nets,” *Theor. Comput. Sci.*, vol. 13, no. 1, pp. 109–135, 1981, doi: 10.1016/0304-3975(81)90113-4.

[577] J. Cheng, C. Liu, M. Zhou, Q. Zeng, and A. Yla-Jaaski, “Automatic Composition of Semantic Web Services Based on Fuzzy Predicate Petri Nets,” *IEEE Trans. Autom. Sci. Eng.*, vol. 12, no. 2, pp. 680–689, Apr. 2015, doi: 10.1109/TASE.2013.2293879.

[578] H. Cheng, L. Yan, Z. Ma, and S. Ribarić, “Fuzzy spatio-temporal ontologies and formal construction based on fuzzy Petri nets,” *Comput. Intell.*, vol. 35, no. 1, pp. 204–239, Feb. 2019, doi: 10.1111/coin.12199.

[579] J. C. Vidal, M. Lama, and A. Bugarín, “OPENET: Ontology-based engine for high-level Petri nets,” *Expert Syst. Appl.*, vol. 37, no. 9, pp. 6493–6509, Sep. 2010, doi: 10.1016/j.eswa.2010.02.136.

[580] A. Pandat and M. Bhise, “RDF Query processing : Relational vs . Graph Approach,” *Futur. Trends Netw. Commun. Technol. FTNCT*, no. November, 2021.

[581] B. Szigeti, Y. D. Roth, J. A. P. Sekar, A. P. Goldberg, S. C. Pochiraju, and J. R. Karr, “A blueprint for human whole-cell modeling,” *Curr. Opin. Syst. Biol.*, vol. 7, pp. 8–15, Feb. 2018, doi: 10.1016/j.coisb.2017.10.005.

[582] and M. L. B. Oliver Ruebenacker, Ion I. Moraru1, “Latest from the Data Integration Frontier: Using Pathway Data to Build and Annotate Systems Biology Models,” vol. 2, no. 2005, p. 2008, 2008, [Online]. Available: https://cnls.lanl.gov/q-bio/wiki/images/0/0b/100_Ruebenacker.pdf

[583] O. Ruebenacker, I. I. Moraru, J. C. Schaff, and M. L. Blinov, “Kinetic Modeling Using BioPAX Ontology,” in *2007 IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2007)*, Nov. 2007, vol. 48, no. Suppl 2, pp. 339–348. doi: 10.1109/BIBM.2007.55.

[584] H. Mi *et al.*, “BioPAX support in CellDesigner,” *Bioinformatics*, vol. 27, no. 24, pp. 3437–3438, Dec. 2011, doi: 10.1093/bioinformatics/btr586.

[585] T. Willemsen, A. Feenstra, and P. Groth, “Building Exceutable Biological Pathway Models Automatically from BioPAX”.

[586] O. Ruebenacker, I. I. Moraru, and M. L. Blinov, “Towards Unifying Systems Biology - Using Pathway Data in Biopax Format for SBML Simulators,” *Biophys. J.*, vol. 96, no. 3, p. 650a, Feb. 2009, doi: 10.1016/j.bpj.2008.12.3432.

[587] Y. Munarko *et al.*, “NLIMED: Natural Language Interface for Model Entity Discovery in Biosimulation Model Repositories,” *Front. Physiol.*, vol. 13, no. February, pp. 1–12, Feb. 2022, doi: 10.3389/fphys.2022.820683.

[588] J. R. Karr *et al.*, “A Whole-Cell Computational Model Predicts Phenotype from

Genotype," *Cell*, vol. 150, no. 2, pp. 389–401, Jul. 2012, doi: 10.1016/j.cell.2012.05.044.

List of publications

N. A. Zondervan, Jesse C. J. van Dam, Peter J. Schaap, Vitor A.P. Martins dos Santos, Maria Suarez-Diez. "Regulation of Three Virulence Strategies of Mycobacterium tuberculosis: A Success Story". In *International Journal of Molecular Sciences* 19(2) 2018. Doi: 10.3390/ijms19020347

Erno, Lindfors*, Jesse C. J. van Dam*, Carolyn Ming Chi Lam, **Niels A. Zondervan**, Vitor A. P. Martins dos Santos, Maria Suarez-Diez. "SyNDI: synchronous network data integration framework". In *BMC Bioinformatics* 19(1) 2018. Doi: https://doi.org/10.1186/s12859-018-2426-5

Niels A. Zondervan, Vitor A. P. Martins dos Santos, Maria Suarez-Diez, Edoardo Saccenti. "Phenotype and multi-omics comparison of *Staphylococcus* and *Streptococcus* uncovers pathogenic traits and predicts zoonotic potential". In *BMC genomics* 22(1) 2021. Doi: https://doi.org/10.1186/s12864-021-07707

Niels A. Zondervan, Vitor A. P. Martins dos Santos, Maria Suarez-Diez. "Predicting *Mycoplasma* tissue and host specificity from genome sequences". Submitted for publication. Preprint available at https://biorxiv.org/cgi/content/short/2022.08.08.503189v1.

Niels A. Zondervan, Eva Yus, Daniel C. Sévin, Sira Martinez, Carolina Gallo, Peter J. Schaap, Maria Lluch-Senar, Luis Serrano, Vitor A. P. Martins dos Santos and Maria Suarez-Diez. Exploring the adaptability and robustness of the central carbon metabolism of *Mycoplasma pneumoniae*. Submitted for publication. Preprint available at https://biorxiv.org/cgi/content/short/2022.08.08.503180v1.

Tolentino-Zondervan, F. & **Niels. A. Zondervan.** *Shared co-authorship. "Sustainable fishery management trends in Philippine fisheries". Ocean and Coastal Management, Vol. 223, 106149. 2022. Doi: https://doi.org/10.1016/j.ocecoaman.2022.106149

Niels A. Zondervan*, Frazen Tolentino-Zondervan*, Dennis Moeke. *Shared co-authorship. "Logistics trends and innovations in response to Covid-19 pandemic: An analysis using text mining". Processes. 2022.

Overview of completed training activities

Discipline-specific activities

Name of the course or meeting	*Organizer*	*City, country*	*Year*
Tutorial Multiscale, Cell-based Modelling in Biological Development and Cancer SRCSB	SRCSB	Stuttgart, Germany	2014
Life Science e-infrastructure Workshop	DTL on track	Utrecht	2014
Practical Integrative Cell Models	Lorentz Center, WUR	Berlin, Germany	2015
BioSB conference	BioSB	Lunteren, The Netherlands	2015
MycoSynVac meeting Barcelona	CRG	Barcelona, Spain	2015
From Big data to biological solutions (Symposium)	WUR	Wageningen, The Netherlands	2015
Brainstorm-symposium on Synthetic Biology	WUR	Wageningen, The Netherlands	2015
BioSB course, Managing and Integrating Life Science Information, Approaches using Linked Data and Semantics	BioSB	Lunteren, The Netherlands	2016
MycoSynVac 1st general meeting	INRAE	Bordeaux, France	2016
MycoSynVac 2end general meeting	MSD	Boxmeer, the Netherlands	2017
Mini Workschop in FAIR DM for WUR	WUR, Corynebacterium project	Wageningen, the Netherlands	2017
Data Management workshop (FairDOM)	MycoSynVac LifeGlimmer	Berlin, Germany	2017
MycoSynVac 3rd review meeting	CRG	Barcelona, Spain	2019
MycoSynVac 2nd review meeting	European Commission	Brussels, Belgium	2018
BIOSB 2018	BioSB	Lunteren, the Netherlands	2018
Mini Workschop in FAIR DM for WUR 2018	WUR	Wageningen, the Netherlands	2018
Dutch Blockchain Coalition - Nationale Blockchain cursus	VU Amsterdam	Online	2020

General courses

Name of the course or meeting	*Organizer*	*City, country*	*Year*
VLAG PhD week	VLAG	Baarlo, The Netherlands	2014
Entrepreneurship in and outside Science	WGS	Wageningen, The Netherlands	2014
Mobilising your - scientific - network	WGS	Wageningen, The Netherlands	2014
Teaching and supervising Theses students	WGS	Wageningen, The Netherlands	2015
Project and time management	WGS	Wageningen, The Netherlands	2017
Scientific writing	WGS	Wageningen, The Netherlands	2017

Optional courses

Name of the course or meeting	*Organizer*	*City, country*	*Year*
Preparation of research proposal	SSB	Wageningen, The Netherlands	2015
SSB retreat	SSB	Hengelo, The Netherlands	2016
PhD lab trip to the west coast US of the US (MIB-SSB lab trip)	MIB & SSB	Wageningen, The Netherlands	2017
Bi-Weekly group meetings	SSB & CSB	Wageningen, The Netherlands	2015-2018
SSB retreat	SSB	Hengelo, The Netherlands	2018

Teaching activities

Name of the course or meeting	*Year*
Molecular Systems Biology	2015
Molecular Systems Biology	2017
Course Metabolic modelling (part of WurSynBio)	2017

Student supervisions

Name of the student	*Year*	*MSC/BSC*
Jianan Chen	2016-2017	Msc
Sardy Partowidjojo	2017-2018	Bsc

Acknowledgements

A PhD is more than a book and a collection of articles. This PhD has been a journey and a life chapter involving many changes. Since I started my PhD, I got engaged, married, bought a house, and was blessed with three amazing kids to enrich my life. **Frazen**, my beloved wife, and my kids **Bastiaan**, **Matthijs** and **Daphne**, the four of you have been my inspiration and my '*raison* d'*être*' during these years. When I think back about this life chapter, without doubt, the four of you have been my greatest 'achievement' and the source of my happiness during these challenging times. This PhD dissertation is dedicated to the four of you! To my parents, **Marian & Koos,** thank you for providing me with a youth filled with books and discussion which helped me to cultivate an open and questioning mind. To my brothers **Gideon**, **Morris**, **& Lars**, thank you for being great brothers, role models, playmates, and discussion partners. To my sister's in-law, **Rohani**, **Francien** and **Marjolein**, thank you for listening to my complains about the hardships of a PhD. To my best friend **Bosko**, you might not be able to attend my defence, but you will be in my thoughts.

Firstly, I would like to thank my supervisors and promotors who made it possible for me to start and complete this PhD journey. I am grateful to you **Peter** for allowing me to explore my vision in the initial phase of my PhD, for being an open-minded idea guy, and for having my back during group and EU meetings while you were my supervisor. I remember that we know each other since my bachelor thesis and that you introduced me to grep, AWK and SED, which still come in handy on a day-to-day basis as white hat hacker and owner of the company Wallet Recovery NL. I am grateful to my promotor **Vitor** for giving me the opportunity to do my PhD thesis at the Systems and Synthetic Biology group, for encouraging me to go to international workshops and for encouraging me to travel to the Philippines in the first year of my PhD to visit Frazen (who is still my girlfriend at that time). You told me good opportunities and good partners are important and worth making time for. **Edoardo**, when I worked on my *Staphylococcus Streptococcus* paper and we co-supervised a student and co-authored a paper, our collaboration was effortless. You are a versatile supervisor. I am most of all grateful to my supervisor and co-promotor **Maria** who I have seen evolve from a Post Doc to a full Professor during my time at the Systems and Synthetic Biology (SSB) group. A position well deserved in my opinion. It might be 'cliche' to hear this, since you have heard it from many PhD's, but without you I would not be standing here today to defend my PhD. In my darkest hours you kept me sane by separating process and output in my supervision. The position of Professor of the SSB group is in good hands with you. I have no doubt that you will lead the SSB group with the same passion and steady hand as you have guided me as well as so many PhD's to the finish line.

I would like to thank my colleagues at SSB and Computational Systems Biology (CSB) group. **Rienk** and **Mark**, our SSB 'elders'. The two of you introduced me to Genome Scale metabolic modelling when I asked you during the course Bioprocess Design, which helped in my decision to perform my master theses at SSB. Rienk, I must especially thank you for helping me structure the jeopardised abundance of information in my master thesis. I published this thesis in a more extended form years

later, of which you should have been involved as a co-author in my opinion, but sadly you were too busy with your job as patent officer. Those early PhD years at the Drije were memorable and I miss your homebrew beers, which you brought to drink with your colleagues after work. **Michael, Milad, Dorett, Ruben, Jasper** and **Jesse.** Who knew that programming a Raspberry Pi candy train could become an actual project! It was fun eating pizzas, chatting while enjoying **Milad's** guitar skills. While most of us were chillaxing, **Michael, Jesse** and **Jasper** hacked away at the Raspberry Pi to pull the project. I enjoyed our times together as much as the magnificent 2-euro Chinese lunches sold from the back of a car at the Drije. Those where the days. After the Raspberry Pi candy train project, I continued with Jasper and Jesse to build our own company, which was a fun and a tremendous educative process. I do however regret that our business ambitions got in the way of pizza nights and outings such as paintball, canoeing and carting. **Ruben**, I enjoyed the Lorentz workshop and the modelling course that we did together in Germany, as well as the whisky and talks we shared during these events and during the annual whisky tasting in the Junushoff. Having you at coffee breaks was great since you were a great indicator when topics got escalated enough to signal it was time to get back go to work. Having a sane mind at the breaks was a good balance for the more 'radical' coffee break members.

Javi, **Karl** and **Benoit**, it was fun sharing an office although the 'chaos' of these good times was a bit challenging for me until I invented working with a headset. Still, I enjoyed our times together and am thankful to you Karl for introducing me to your home-made curies. **Benoit** you saved me more than once from computer disaster. For example, when I had the brilliant idea of changing my Linux installations by accident to only recognized my remote monitor at home, you manage to repair my completely broken system. Thanks for being our Linux guru together with **Bart**. **Bart** and **Nicholas**, sharing an office with the two of you was fun. The relative calm in our room was great for my productivity. **Benoit,** you really embody the work hard, party hard ethos. Although you often showed up late in office and appeared easy going, you worked hard till late hours and still found time to help your colleagues with their Linux issues. Having fellow Otaku's in the same office room was a pleasant surprise. **Bart,** apart from being a fellow Otaku, it was great to get to know you and your family as I learned we shared more than our passion for anime and computers, such as similar ideas about living and child raising**. Nicholas**, who knew my beloved Kalamata olives, which I highly praise in our office, were the pride of your home region in Greece**.** Do bring me some when you visit your hometown. **Maarten,** my favourite 'bioinformagician' and fellow gamer. If I would have known you a few years earlier before I became 'burgerlijk' (married with kids), I am certain we would have enjoyed many more LAN parties and beers together while being the opposite of burgerlijk. Together with our favourite SSB member, **Bastian**, you knew how to systematically entice people to have regular coffee breaks. **Bastian** and **Maarten** also know how to throw the best parties. **Niru** and **Emma**, thank you for keeping **Maarten** and **Rob**'s fun but sometimes offensive jokes in check. Thank you two for laughing when I made double meaning hidden green jokes. Your laughter revealed both your true nature as well as your great intelligence;) **Nong, Nhung, Marta** and **Linde,** most of you joined later in the group but soon you became the new social coherence factors after our

beloved **Basti** finished his PhD. Thank you for 'restoring balance to the force' by reducing the nerd factor in our group and bringing in fresh enthusiasm and social spirit. **Wasin,** thank you for breaking my prejudice against Asian men by showing me that they can be light and have a great sense of humour. **Stamatios**, it was always a guess whether you were in the office, the lab or traveling to Greece. If I would give you a nickname, it would be 'doctor plasmid' since your enthusiasm for these vectors knows no limit. **Erika,** and **Tjerko,** having the two of you involved in the same project was great. **Tjerko,** thank you for entertaining my *Mycoplasma* theories on metabolism and for introducing your file naming convention of *yyyy-mm-dd* to the group. I used this file system since both for personal and professional use. **Erika,** next to a common interest in *Mycoplasma* we shared a passion for Bordeaux wines, Grand Cru, and oysters. It was great sharing them with you as well as listening you gloat when describing how the Dutch fail to make even the simplest Italian pasta meals. **Anna**, you only joined our office later. It was a great to have another parent to talk to in the office. I realise some of the challenges in life and in doing a PhD while having kids are only understood by fellow parents. Thank you for shuttling me to the hospital when I was in need and for brightening our office. **Sara**, **Maria,** and **Christos** thank you for keeping the office a welcoming place after my contract ended. **Jesse** and **Jasper**, having regular break walks with you was both fun, inspirational, and necessary. It was the only thing that kept me somewhat sane when I was sleep deprived when our first baby was born and while I was facing challenges in my PhD project. I think we both learned a lot during our journey to build our company Big Bio Data (Simini) together. More importantly, we became friends for life. To the many others of you such as **Sanjeevan, Rita, Enrique, Wen, Yuan, Alex, Rik, Anna, Sven, Irene, Peer,** and those I might not have specifically mentioned here, thank you for making my time at the Systems and Synthetic Biology (SSB) group and the Computation Systems Biology (CSB) group fun and memorable.

Funding

The research described in this thesis was financially supported by the European Union through the FP7 programme under grant agreement No. 305340 (INFECT), the SysteMTb project (HEALTH-F4-2010-241587) and the Horizon 2020 research and innovation programme under grant agreement No. 634942 (MycoSynVac).

Cover design by: Niels Zondervan

Printed by: Proefschriftmaken.nl on FSC-certified paper