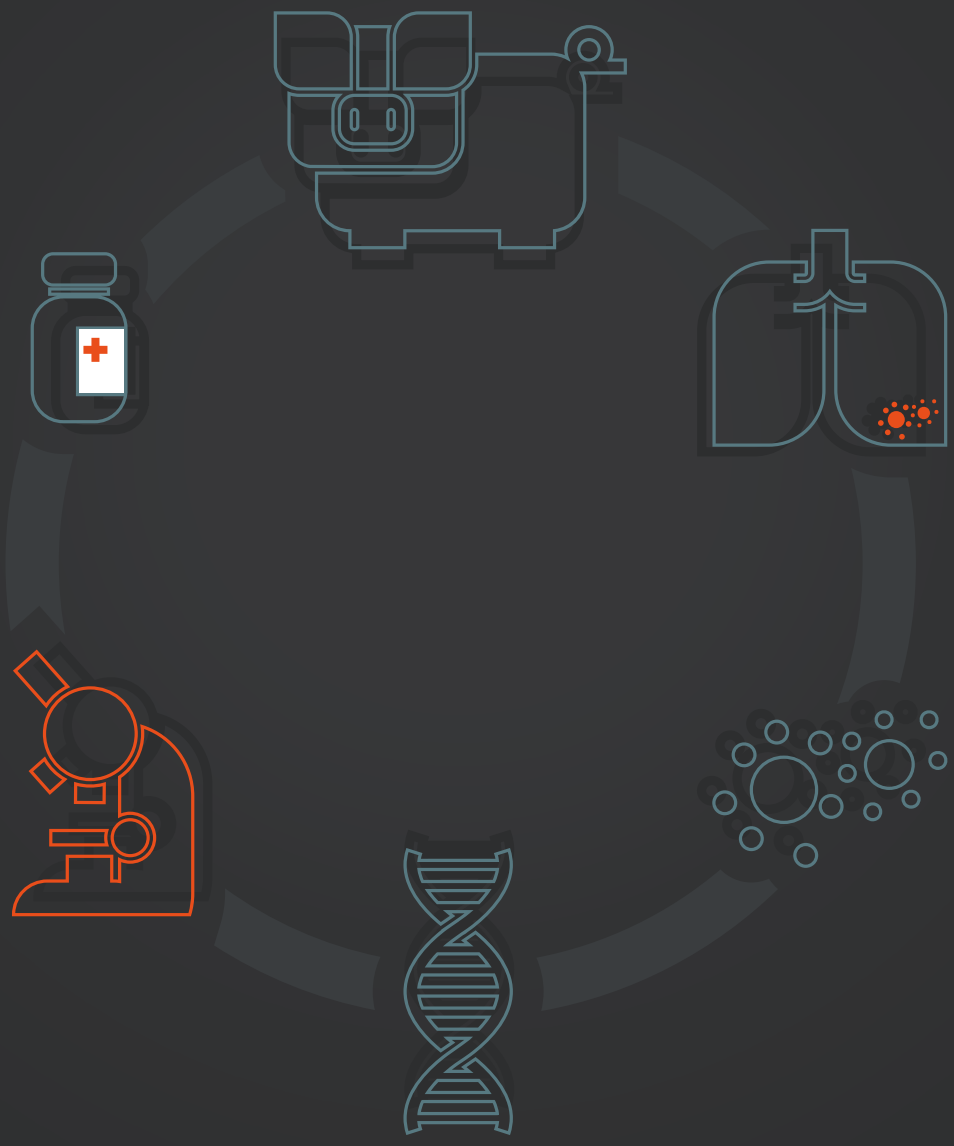
Systems analysis of *Mycoplasma hyopneumoniae* to improve vaccine production



Tjerko Kamminga

Propositions

1. To maintain cellular homeostasis, 84% of cellular energy is used for non-growth associated maintenance in a standard *M. hyopneumoniae* aerobic fermentation.

   (this thesis)

2. Genome-scale metabolic models reflect our inability to understand even the simplest self-replicating organisms.

   (this thesis)

3. Obesity is in principle a flux balance problem which should allow dietitians to simulate personal diets using genome-scale metabolic models.

4. Design of personalized medicine based on social media input is prone to bias.

5. The renewable energy revolution will be consumer-driven.

6. Ability to control scope creep is an essential skill for a scientist.

7. Adapting to stationary growth will be one of the major challenges for future human generations.

Propositions belonging to the thesis, entitled:

'Systems analysis of *Mycoplasma hyopneumoniae* to improve vaccine production'.

Tjerko Kamminga

Wageningen, 16 November 2017.

Systems analysis of *Mycoplasma hyopneumoniae* to improve vaccine production

Tjerko Kamminga

Thesis committee

Promotor
Prof. Dr V.A.P. Martins dos Santos
Professor of Systems and Synthetic Biology
Wageningen University & Research

Co-promotors
Dr P.J. Schaap
Associate professor, Systems and Synthetic Biology
Wageningen University & Research

Dr J.J.E. Bijlsma
Associate Principal Scientist, Discovery & Technology
MSD Animal Health, Boxmeer

Other members
Prof. Dr H.F.J. Savelkoul, Wageningen University & Research
Prof. Dr B. Teusink, VU Amsterdam
Prof. Dr M.H.M. Eppink, Wageningen University & Research
Dr F. Tardy, ANSES, Laboratoire de Lyon, Lyon, France

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

Systems analysis of *Mycoplasma hyopneumoniae* to improve vaccine production

Tjerko Kamminga

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 Wednesday 16 November 2017
at 11 a.m. in the Aula.

Tjerko Kamminga
Systems analysis of *Mycoplasma hyopneumoniae* to improve vaccine production, 162 pages

PhD thesis, Wageningen University, Wageningen, The Netherlands (2017)
With references, with summaries in English and Dutch
ISBN: 978-94-6343-651-9
DOI: 10.18174/420702

Contents

Chapter 1

General introduction

Porcine respiratory disease is arguably the largest worldwide threat for pig herd health. The etiology of disease varies depending on herd size, environmental conditions, presence of infectious agents and herd management practices. Major contributors to the development of disease in the later stages of pig production are porcine circovirus type 2 (PCV2), porcine reproductive and respiratory syndrome virus (PRRSV) and *Mycoplasma hyopneumoniae*[1–3].

Mycoplasma hyopneumoniae disease

Mycoplasma hyopneumoniae (*M. hyopneumoniae*) is a bacterial pathogen that causes enzootic pneumonia in pigs. Disease symptoms are mild but morbidity is high and presence of this pathogen in a pig herd often causes secondary infections, which makes it an important contributor to development of respiratory disease. Pigs infected with *M. hyopneumoniae* often develop a dry, non-productive cough, gain less weight and, when slaughtered, show lung lesions. Pigs can be infected by other pigs carrying the disease via nose-to-nose contact (e.g. sow to piglet), coughing and empirical evidence has shown that infection can occur via aerosols over long distances[4]. The persistent nature of the disease, the high infectivity and the easy transmission makes it difficult to maintain a *M. hyopneumoniae* free herd.

Treatment and prevention

Severity of the disease can be controlled with antibiotics (tylosin, lincomycin, tiamulin or tetracycline), by improvement of housing conditions, or herd management practices[5]. Treatment with antibiotics in general leads to a decrease in clinical signs and the amount of lung lesions but does not eradicate the disease. Applying an all-in all-out production system was shown to be effective because transmission from older to younger pigs was prevented and there is less stress caused by mixing and sorting of different pig populations. Crowding of facilities could lead to increased transmission and therefore a decreased animal density could reduce the level of disease[5]. A correlation between herd size and severity of the disease has not been established but optimization of housing conditions (e.g. temperature and ventilation) is important to control infections[6].

Vaccines

Over 75% of piglets in large breeding herds and a majority of pigs in nursery phase or grower/finisher phase were vaccinated against *M. hyopneumoniae* in the US in 2012[3], using inactivated whole-cell vaccines. These vaccines can be applied intradermally or intramuscularly and recently combination vaccines have been developed which protect against *M. hyopneumoniae* and PCV2. Efficacy of these vaccines was tested in vaccination-

challenge studies which showed that vaccination against *M. hyopneumoniae* reduced clinical signs, reduced lung lesions, lowered medical treatment cost, improved daily weight gain and increased the food conversion ratio[7,8]. The exact working mechanisms of these vaccines are not yet understood and it is not known which components in the inactivated whole cell antigen elicit protective immunity. It is expected that both mucosal antibodies and cell-mediated immunity are required[9]. Vaccination was shown to increase *M. hyopneumoniae* specific antibody levels in serum and in the respiratory tract and induced IFN-γ secretion in blood. However, there is no clear correlation between the local presence of *M. hyopneumoniae* specific antibodies and protection[10,11]. Furthermore, systemic cell-mediated immune responses, measured by proliferation of lymphocytes, were shown to be highly variable amongst different pigs in a study group[10]. Additionally, in vaccination-challenge studies it was shown that vaccinated-challenged pigs had lower levels of TNF-α in bronchial alveolar lavage fluid[9] and reduced macrophage infiltration in bronchus-associated lymphoid tissue[12]. Many researchers have investigated sub-unit, DNA or recombinant vector vaccines against *M. hyopneumoniae*[13] but efficacy of these vaccines in pigs was found to be insufficient.

M. hyopneumoniae phylogeny and genomics

M. hyopneumoniae is present in the Hominis phylogenetic cluster within the mycoplasma genus. Mycoplasma species have evolved by genome reduction from a gram-positive ancestor and have adapted to growth in a specific vertebrate host. As a result of reductive evolution, the metabolic capabilities of mycoplasma species were minimized and there is a high dependence on the host to supply nutrients needed for growth. Major metabolic pathways that are lacking are the citric acid cycle, oxidative phosphorylation and the oxidative part of the pentose phosphate pathway (PPP). Furthermore, there are no annotated enzymes for reactions related to amino acid synthesis, synthesis of nucleotide bases, synthesis of fatty acids, synthesis of cholesterol or synthesis of vitamins[14]. Mycoplasma species have no cell wall and two-third of the mass of the cellular membrane consists of proteins, e.g. transporters and lipoproteins[14]. Currently, the genomes of six *M. hyopneumoniae* strains have been fully sequenced (232, J, 7422, 7448, 168 and 168-L). The first strain sequenced was strain 232, which had a genome-size of 892.759 basepairs, contained 692 predicted protein coding sequences and a GC content of 28.6 mol%[15]. The function of 56% of the predicted protein coding sequences was unknown. Most sequenced strains are pathogenic but strains J and 168-L are non-pathogenic[16,17]. Strain pathogenicity was determined in challenge studies in pigs where it was found that pigs challenged with strain J cultures showed no lesions[18] and in pigs challenged with strain 168-L no ciliary damage was observed[17]. In addition, strain J had a reduced ability to adhere to cell monolayers[19]. Comparative genomics studies between pathogenic and non-pathogenic strains have identified multiple differences in genome regions encoding putative virulence

factors[16,17], however, a single factor causing strain attenuation was not found. To trick the immune system, mycoplasma species genomes contain tandem repeats[20] in genes encoding surface proteins, especially lipoproteins, where slipped strand mispairing occurs during DNA replication resulting in variations in gene length or phase switching. However, the amount of repeat sequences found in the genome of *M. hyopneumoniae* is relatively low compared to other mycoplasma species[15] and it is not clear how *M. hyopneumoniae* evades the host immune response and causes chronic infections.

M. hyopneumoniae infection mechanisms

Immunofluorescence and electron microscopy studies have shown that *M. hyopneumoniae* mainly colonizes the lower respiratory tract[21] (trachea, bronchi and bronchioles) through attachment to cilia on the luminal surface of the respiratory epithelium (figure 1). Low numbers of *M. hyopneumoniae* were also identified in nose swabs[22] using PCR and few were present in alveoli[23]. The mechanisms used to attach to the cilia are well-described. The *M. hyopneumoniae* genome contains a family of adhesion proteins which bind to glycosaminoglycans (e.g. heparin), fibronectin or plasminogen. P97 was the first protein shown to bind heparin and is referred to as the cilium adhesin[24]. The C-terminal end of P97 contains two repeat regions of which the lengths vary between *M. hyopneumoniae* genomes[25]. The R1 region contains the repeat AAKPV/E[26] of which a minimum of 8 copies is needed for binding to the cilia. P97 is present in an operon together with P102, which binds fibronectin and plasminogen[27], both genes have 6 paralogs in the *M. hyopneumoniae* strain 232 genome[15]. These adhesins are proteolytically cleaved[27–32] with varying efficiency which allows the organism to vary the composition of the membrane surface proteins. Besides the P97/P102 paralogous families, two other *M. hyopneumoniae* genes were described to function as cilium adhesins: P159 (MHP_RS02540) is a cilium adhesin[33] which is cleaved into three fragments that bind heparin and MHP_RS01270 which is present on the cell surface and binds heparin and plasminogen[34]. Binding of *M. hyopneumoniae* to the ciliated epithelium could cause ciliostasis, loss of cilia, cilia clumping and death of epithelial cells[35]. Virulence factors responsible for damaging the cilia and epithelium are not yet fully known. It is possible that production of hydrogen peroxide by glycerol oxidase plays a role[36]. Furthermore, recruitment of plasminogen and conversion to plasmin by the host could cause damage to the host extracellular matrix proteins[27,34,37,38]. Lipoproteins expressed on the bacterial surface have attracted much attention because they were found to be highly immunogenic[39] but their role during infection is unknown. The exact role of the adhesive proteins and proteins possibly related to virulence still has to be verified through the generation of gene knock-out mutants. Part of the damage to lung tissue could be caused by the action of pro-inflammatory cytokines at the site of infection. *M. hyopneumoniae* infection attracts macrophages, neutrophils and lymphocytes. Histologically, massive peribronchiolar, peribronchial and cardiovascular

lymphoid hyperplasia was observed[40]. Examination of infected bronchus-associated lymphoid tissue with immunohistochemical methods showed presence of TNF-α, IL-1, IL-4, IL-6, IL-8 and prostaglandin E_2[41–44]. Finally, the *M. hyopneumoniae* encodes several nucleases[45], proteases[46] and lipases[47] and a close association or even fusion between mycoplasma species and host cells could cause damage to the host cells as a result of activity of these enzymes[48]. The interactions between *M. hyopneumoniae* and the host should be further investigated to understand the contribution of the various pathogenic mechanisms to development of disease.

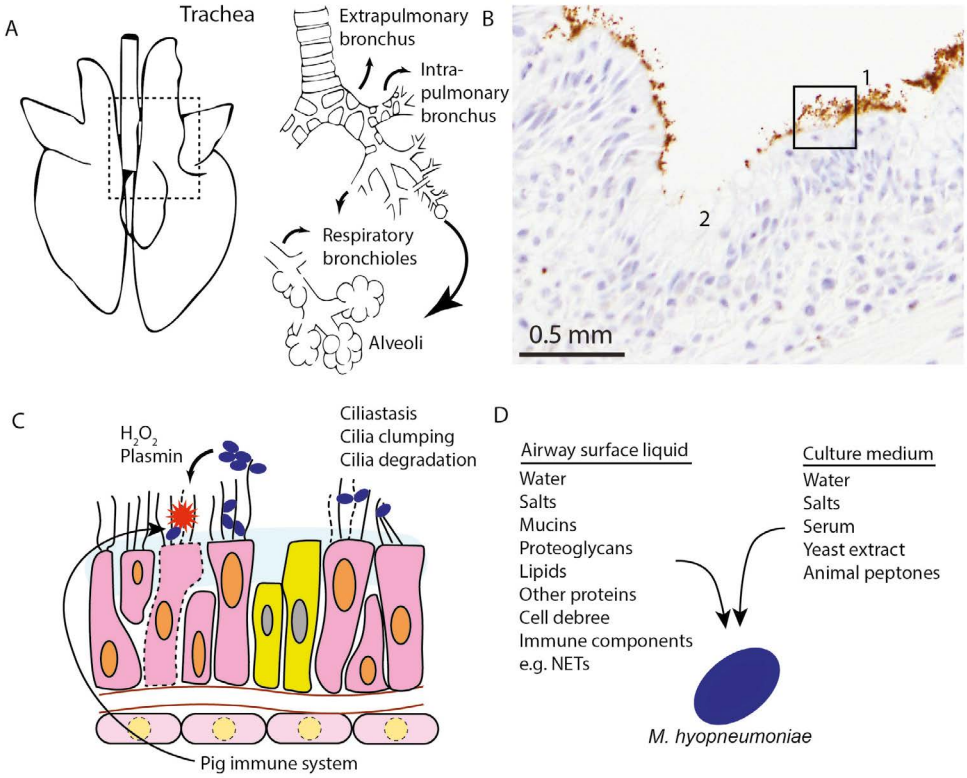


Figure 1. ***Mycoplasma hyopneumoniae* specifically binds to the ciliated respiratory epithelium.** A: Pig lung anatomy, *M. hyopneumoniae* resides on the ciliated epithelium of the trachea, extrapulmonary bronchi, intrapulmonary bronchi and respiratory bronchioles. Pictures adapted from Frandson *et al.*[49] and veterinaria.org[50]. B: Immunohistochemistically coloured intrapulmonary bronchus of an *M. hyopneumoniae* infected porcine lung. Immunolabelling of protein P46 of *M. hyopneumoniae* shows presence of bacteria on the cilia of the respiratory epithelium (1). Goblet cells are indicated (2) which produce mucus, a possible source of nutrients for the bacterium. C: Schematic drawing of the interaction between *M. hyopneumoniae* (blue elipses) and the ciliated epithelium (pink). *M.* hyopneumoniae causes ciliostasis, clumping of cilia and cilia degradation. Expected pathogenic mechanisms are shown: production of hydrogen peroxide and plasmin, additionally tissue damage is expected due to the response of the pig immune system. Goblet cells are shown in yellow. D: Possible nutrients available for growth in airway surface liquid and culture medium.

M. hyopneumoniae transcriptomics

Transcription regulatory mechanisms in *M. hyopneumoniae* are poorly understood. There is only a single sigma factor present in the *M. hyopneumoniae* genome, no transcription termination Rho factor is found and few regulatory proteins[51]. Promoter sequences have been determined by comparative analysis of transcription start sites and a -10 element was found but a -35 element could not be distinguished[52]. Transcriptional termination is expected to occur via the formation of stem-loop structures[53]. The majority of genes expressed in *M. hyopneumoniae* are transcribed as polycistronic mRNAs forming transcription units that often show stair-case behavior in their transcription levels, meaning that there is a decrease in expression level depending on the distance from the transcriptional start site[54–56]. A possible role for tandem repeats and palindromic elements in the 5' region upstream of transcription units has been described[57]. Transcriptional landscapes in *M. hyopneumoniae* should be further investigated to gain insight in the significance of these transcriptional mechanisms.

Production process challenges

Due to the fastidious growth requirements, complex components are needed in the cultivation medium used for production of whole-cell inactivated vaccines, such as porcine serum and peptones obtained from swine or bovine tissue. Use of these complex components is unwanted because the chemical composition is non-defined and inherently variable as a result of the biological origin of the starting materials. Furthermore, there is a risk that extraneous agents are present in these components which requires additional inactivation steps before these components can be used. The presence of serum protein in the medium also poses a challenge for the filtration step in downstream processing because it is hard to separate serum protein from biomass. Finally, quality control of whole-cell inactivated *M. hyopneumoniae* antigen is complicated because there is no single known virulence factor that correlates with protection. Solutions to these problems should be found by better understanding the growth requirements and host-pathogen interactions during infection.

System Biology to tackle production process challenges

Systems biology aims to understand complex biological systems by breaking them down into their essential components and identifying the dynamic interactions between these components both chemically and physically. A systems analysis of *M. hyopneumoniae* requires studies on multiple cellular levels (figure 2). In this thesis project, we studied *M. hyopneumoniae* functional capability based on the protein domain content of the genome,

analyzed metabolic capability using a genome-scale metabolic model and applied transcriptomics to study gene expression in the host. Here these concepts and the methods used will be shortly introduced.

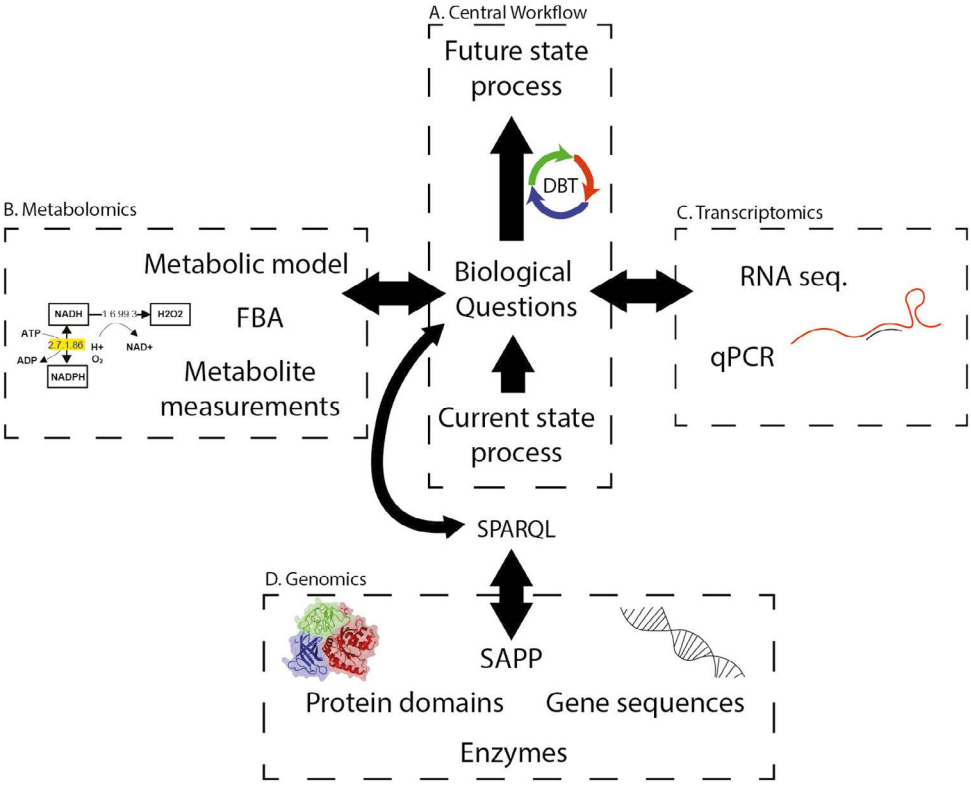


Figure 2. **A systems framework to improve vaccine production.** A: Aim of the thesis is to use systems tools to improve the production process for *M. hyopneumoniae* vaccines. B: Biological questions related to metabolic capability are addressed using metabolic modeling, flux balance analysis and metabolite measurement. C: RNA sequencing and qPCR provides the inputs for transcriptome questions. D: Datawarehouse: Genome specifics are stored in a semantic RDF database and retrieved using SPARQL queries. The database contains genomics information with provenance derived using SAPP.

Domain-based annotation of protein function

Proteins are the working machinery of a cell and functionality of a protein is determined by its protein domain content. Protein domains are well-defined 3D-structures in proteins which perform a specific function in an organism. Examples are: substrate binding regions, helix-turn-helix motifs or phosphate binding structures. Proteins can contain multiple protein domains which operate sequentially of simultaneously to perform the protein function. Based on the sequence of a protein coding region in the genome, multiple algorithms can be used to annotate protein domains[58]. The complete repertoire of protein

domains in an organism, the domainome, describes the functional capability of an organism and provides the basis for functional comparison of bacteria as well as the basis for further systems analysis. As algorithms are used to obtain functional predictions automatically, measures have to be taken to warrant reproducibility and interoperability. We used the semantic annotation pipeline with provenance (SAPP[59]) for genome annotation where data is stored into an RDF (Resource Description Framework) data model. For gene annotation it is important to know that mycoplasma species use codon table four for protein translation, which means that the UGA codon encodes tryptophan instead of a stop codon. SPARQL queries (figure 2) are used to derive information from an RDF database which could be: genes or proteins with a metabolic function, genes or proteins with a specific protein domain or all genes related to a specific cellular function. For the purpose of interoperability, throughout this thesis the RDF data model was used to store biological data with data provenance in a graph database.

Genome-scale metabolic models

Biological questions related to metabolic capability of *M. hyopneumoniae* can be addressed using genome-scale metabolic models. Genome-scale metabolic models have become indispensable for quantitative understanding of metabolic flux distributions during bacterial growth and provide the platform for model-based improvement of the cultivation step for vaccine production (figure 3). Metabolic networks can be represented by a mathematical model describing the stoichiometry of reactions and the respective flux (turnover rate, often defined in $mmol \cdot gDW^{-1} \cdot h^{-1}$) through each reaction[60,61]. An initial metabolic network representation can be created automatically based on the genome annotation of a bacterial strain by retrieving genes with an enzymatic or transportation function[62]. Often, these automatically created networks are incomplete and missing reactions can be assigned based on the structure of the metabolic network by gap-filling algorithms[63] and by manually curating the network based on expert knowledge, experimental data and scientific literature[64]. When reactions are added without an annotated function in the genome, such a reaction is named an orphan reaction. In the mathematical model of metabolism, reaction stoichiometries are defined in a stoichiometric matrix with m rows representing the metabolites and n columns representing the reactions in the metabolic network[60]. Each column in the matrix contains the stoichiometric coefficients for metabolites participating in the reactions. These reaction stoichiometries put constraints on the model solution space, ultimately dictated by the need to balance mass in the system. One of the primary applications for genome-scale metabolic models is the calculation of network flux distributions under varying conditions. Calculating a flux distribution through a network requires flux balance analysis and for biologically meaningful interpretations some fluxes through the model will need to be constrained.

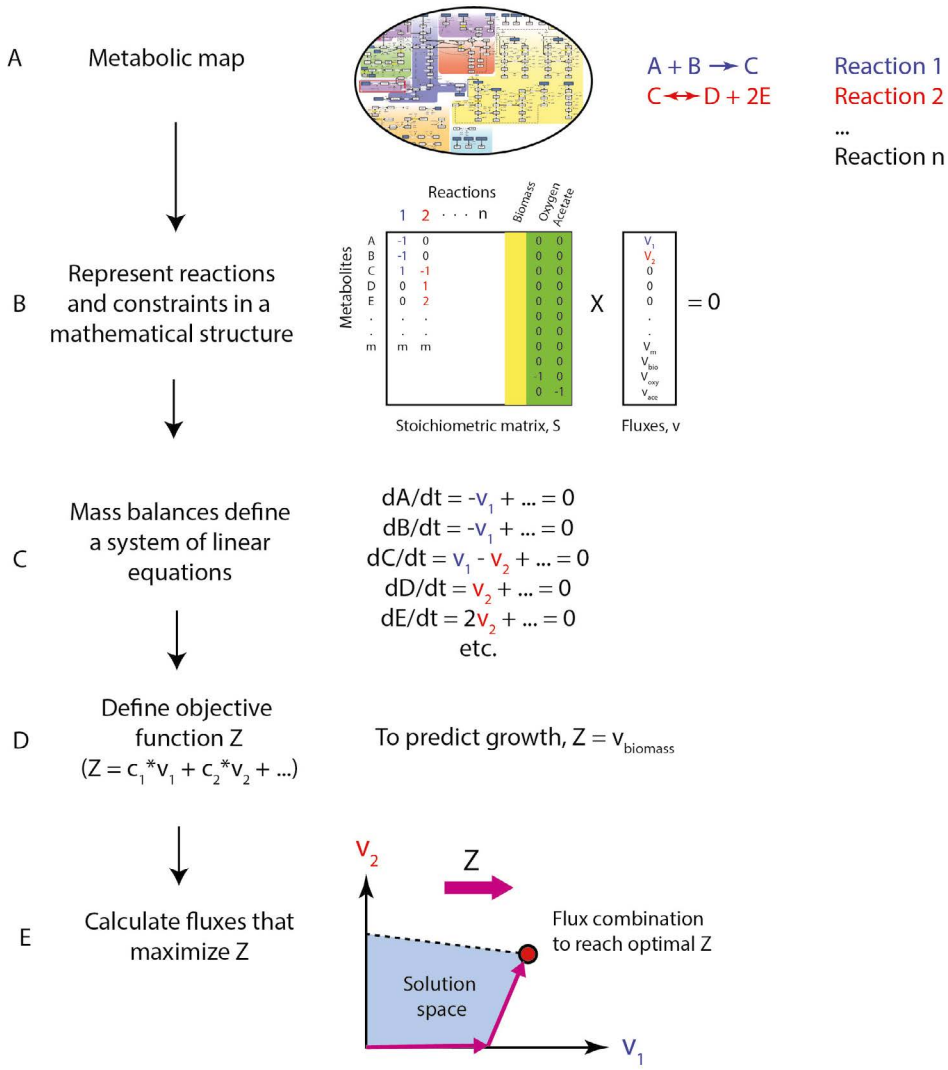


Figure 3. **Flux balance analysis to calculate maximal *M. hyopneumoniae* growth rate.** Steps needed to define a flux balance analysis problem are shown and calculate an optimal solution using linear problem solving. A: Chemical reactions are obtained from the metabolic map created based on the genome annotation and scientific literature. B: Reactions are represented in a stoichiometric matrix where the rows represent metabolites, columns represent reactions and numbers in the matric represent reaction stoichiometries. General equations are added for biomass formation (yellow) and exchange reactions (green) which represent flux into or from the extracellular compartment. C: For each metabolite a mass balance equation is defined forming a set of linear equations that can be solved when it is assumed that the system is in steady-state. D: The objective function is defined, for instance the rate of biomass formation. The weighing factor (c) represents the contribution of a reaction rate to the objective function. E: Linear problem solving is used to find the optimal solution for Z taking into account the constraints set for the solution space (reaction stoichiometries, maximum or minimum fluxed and directionalities). Figure adapted from Orth *et al.*[60]

For instance, the maximum rate of glucose uptake per unit biomass, which depends on the concentration of transporters on the membrane, the efficiency of transport and possibly allosteric control will need to be constrained to represent a biological meaningful value. Another feature of the network that should be determined is the reversibility of the reactions in the network. When the network structure is determined and constraints are set, linear problem solving can be used to find the maximal or minimal flux through a reaction under steady-state assumption. Steady-state assumption means that the change in the concentration of each metabolite in the model in time is zero. For metabolites obtained from the medium or produced as by-products of metabolism the same requirement holds and therefore source and sink reactions are added to supply nutrients or remove by-products. When an equation for biomass is added to the model, the maximal growth rate can be calculated *in silico*. Flux balance analyses using constraint-based metabolic models have been applied in numerous metabolic engineering studies to rationally design bacterial strains that overexpress metabolic components[65,66]. For mycoplasma species, metabolic models have been created and were applied to study growth medium compositions and intracellular energy balances[67–69].

Transcriptome analysis

There is a multitude of analytical methods that can be used to determine the concentration of RNA transcripts in a bacterium. In this thesis qRT-PCR (quantitative Reverse Transcriptase-Polymerase Chain Reaction) and RNA sequencing were used. Both methods are well-established but are based on different principles. qRT-PCR requires a specific primer set used to amplify a target sequence using PCR. The product concentration profile can be measured in time and depends on the start concentration of the template containing the target sequence. RNA sequencing can be used to globally quantify all RNA in a bacterium which is done by determination of the sequence of small RNA reads. Reads can be mapped computationally onto the genome sequence and the number of reads per gene represents expression levels. Both, qRT-PCR as well as RNA sequencing provide answers to biological questions related to gene expression in *M. hyopneumoniae* (figure 2).

State of Art at the start of the PhD project

There is a high interest in the scientific community for bacteria with minimized genomes. This is driven by both fundamental as well as practical reasons. Knowledge of the minimal essential gene set needed to support cellular replication without a host will be a great scientific accomplishment. Furthermore, knowledge of this minimal gene set will provide genetic engineers with a completely defined cellular platform to study cellular functions and test genetic circuits. The genomes of mycoplasma species are the smallest known genomes capable of supporting autonomous cellular replication and are therefore an as close as possible approximation to minimal essential life that nature can provide. Some

groups investigated the evolutionary process of genome reduction[70–72], other groups studied mycoplasma related diseases, but a majority of research efforts were focused on understanding which components are needed for the design of minimal cells. This led to a large body of scientific publications around the period when this thesis project started, focusing on elucidation of e.g. gene essentiality[73], transcriptional mechanisms[55], proteome characteristics[74] and the elucidation of metabolic networks in mycoplasma species[75]. However, these studies mainly focused on the human pathogens *Mycoplasma pneumoniae* and *Mycoplasma genitalium* which are in a different phylogenetic group compared to *M. hyopneumoniae* so findings from these studies could not be directly translated to *M. hyopneumoniae*. A wealth of information was available for *M. hyopneumoniae* regarding the cilium adhesive proteins and their role in host interaction. It was also already early identified that lipoproteins were important immunogenic components on the cell membrane but that these components did not offer sufficient protection as a subunit vaccine[13]. Multiple genomes were sequenced for *M. hyopneumoniae*, and from the genome sequences an estimation of the metabolic capabilities was derived[15,16], but for vaccine production complex media were still used which resulted in unstable yields in the production process for bacterin vaccines.

Thesis outline

The overall aim of this thesis was to understand growth and survival strategies of *M. hyopneumoniae* during infection, to integrate this knowledge with metabolic modeling under conditions used for vaccine production and apply this knowledge to improve the current production process for *M. hyopneumoniae* vaccines.

We first investigated how bioengineering techniques could best be applied to improve a production process for complex bacterial antigens (**chapter 2**). Therefore, we sequenced the *M. hyopneumoniae* production strain and, based on the annotated functions in the genome, we created a manually curated metabolic model, which was subsequently used for dynamic modeling of the fermentation step in the vaccine production process (**chapter 3**). The metabolic model helped to improve yield of the production process but could not be applied to predict a defined medium. To better understand why *M. hyopneumoniae* specifically colonizes the pig and is dependent on components from porcine origin, we analyzed the functional capabilities of 80 mycoplasma species in relation to their host and niche (**chapter 4**). This analysis allowed us to pin-point protein domains possibly important for growth in the pig but whether these proteins played a role during infection was not clear. Therefore, in **chapter 5** we determined the *in vivo* transcriptome using RNA sequencing and compared the *in vivo* transcriptome to the transcriptome in broth grown cultures. We found up-regulation of the F1-like ATPase, nucleases, spermidine and glycerol-3-phosphate transporters and down-regulation of genes related to cilium adhesion (P102), cell division and glycerol uptake. In this study, we also paid attention to the role of

possible regulatory small RNAs as these were recently found to play an important role during infection in other bacterial pathogens. We used our mycoplasma basis for a global study involving multiple bacterial pathogens and found an exponential relationship between the AT content of genomes and the number of ncRNAs transcribed (**chapter 6**). In the final chapter of this thesis (**chapter 7**) I discuss four novel system strategies possibly used in *M. hyopneumoniae* and explain how these can be incorporated in the genome-scale metabolic model and provide an outlook for future studies to further apply systems biology tools to improve the vaccine production process.

Chapter 2

Risk-based bioengineering strategies for reliable bacterial vaccine production

Tjerko Kamminga, Simen-Jan Slagman, Vitor A.P. Martins dos Santos, Jetta J.E. Bijlsma and Peter J. Schaap

Submitted for publication

Abstract

Reliable production of bacterial vaccines necessitates robust production strains and optimized process controls in a conjoint process to reach optimal antigenic mass yields. While bioengineering techniques have been successfully applied to improve strains for production of biochemicals, rational design and subsequent bioengineering of production strains for bacterial vaccines has been hampered by an incomplete understanding of critical process parameters and a lack of early focus on risk-based process development. Early identification of process risks, captured in a description of critical process parameters, is required for optimal vaccine bioengineering. Here, we propose to use a suite of system metabolic engineering tools, integrated in the vaccine development pipeline, to rationally design production strains and conditions for bacterial vaccines. Model-based understanding of the relation between metabolic flux and production of antigenic mass forms the basis for our approach. In addition, we propose to implement process analytical techniques emanating from model-based interpretation of high-throughput (omics) data to monitor and control the production process. Implementation of this workflow requires collaborative process risk assessments performed by academia and industry at multiple stages in the project.

Introduction

We are yet beginning to understand the multi-level interactions between bacterial pathogens and their hosts but, fortunately, this lack of knowledge has not stopped researchers to develop effective vaccines for humans and animals. Due to global initiatives such as the WHO global vaccine action plan[76] to make vaccines more accessible for the human population and initiatives to reduce the risks of unchecked antibiotic use in livestock, demands for new, better and cheaper vaccines are rising, fueling the need to increase the reliability of the production processes for bacterial vaccines. Five types of antigen: whole-cell bacteria (live-attenuated or inactivated), subunits, toxoids and polysaccharide conjugates (figure 1), are used in bacterial vaccines.

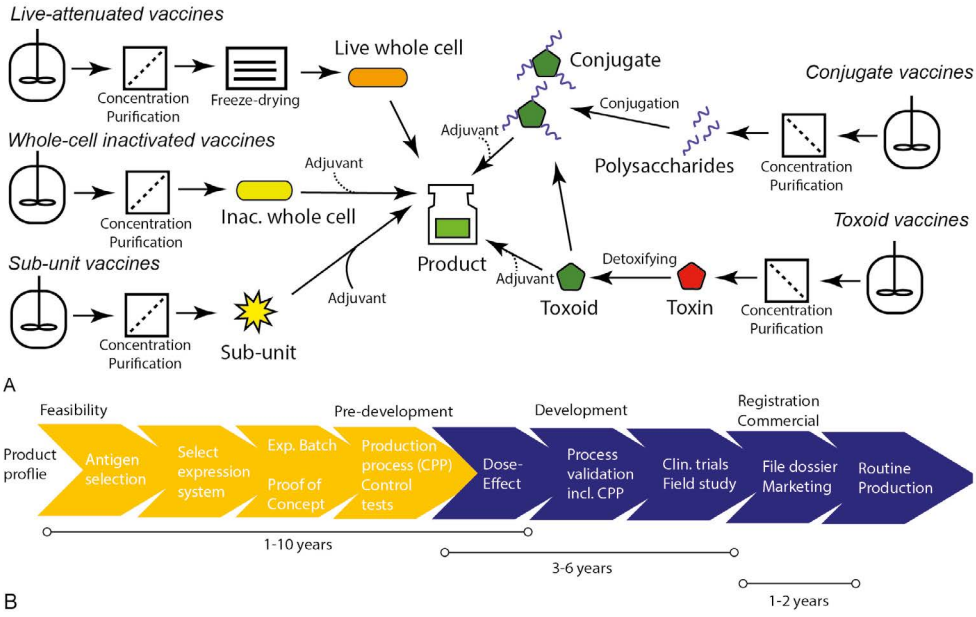


Figure 1. **Production platforms for bacterial vaccine production and the vaccine development pipeline.** A: Simplified platforms for production of bacterial antigens. Each process starts with cultivation of bacteria in controlled fermentor systems to reach high biomass and/or antigenic mass concentrations. Cultures are subsequently concentrated and purified using a variety of methods[77] (e.g. centrifugation, precipitation or filtration). After purification an adjuvant could be added to increase the immune response. B: Stages in a generalized vaccine development process and the final result: a reliable production process. Industry average timelines are shown for separate development stages. Bioengineering techniques should be applied in the first four steps in the vaccine development pipeline (indicated in orange).

Development and introduction of new vaccines is a costly and time-consuming process that is influenced by multiple consequential and long-term factors, including affordability, availability, and safety. A vaccine development process takes on average 8-16 years and

consists of four main stages: i) feasibility, ii) pre-development, iii) development and iv) registration/commercialization[78] (figure 1). Protective antigens used in bacterial vaccines are often undefined, necessitating the use of whole cell vaccines (figure 1, whole-cell inactivated and life-attenuated vaccines). In addition to the need for complex antigens, selected strains are often not optimized for growth in a bioprocess environment and the cultivation media used are often not chemically defined. These major drawbacks have a negative impact on process reliability and affect all bacterial vaccines currently on the market (figure 2).

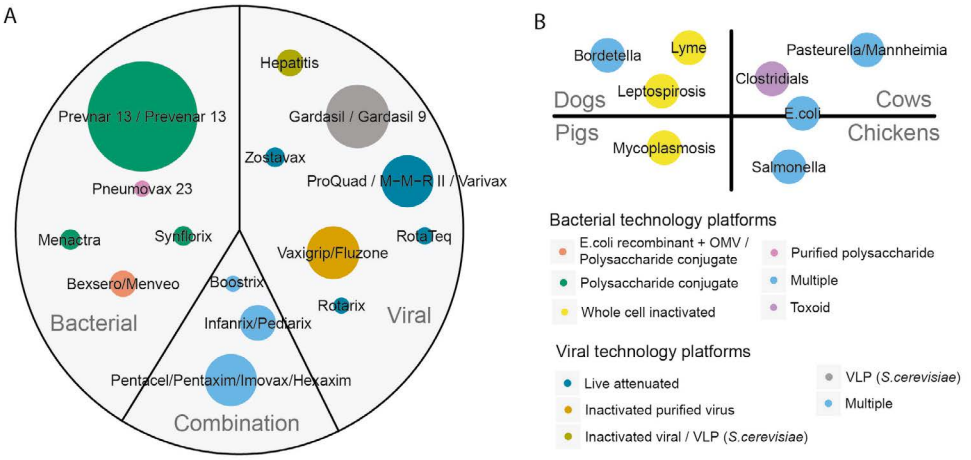


Figure 2. **Overview of the human and animal vaccine market divided by antigen type and technology platforms used.** A: The global human vaccine market was worth over 38.5 billion US$ in 2015[79] divided over bacterial, viral and combination products. Top 15 vaccines are shown according to product sales in 2016, colours indicate manufacturing technology, circle diameter represents sales (1-5 billion US$). B: Major bacterial antigens used in animal vaccines in single and combination products. Colours indicate manufacturing technology; the circles are not scaled for sales.

Already early in the feasibility stage researchers need to lay the foundation for a reliable production process but are depending on complex and variable inputs, non-optimal strains and complex outputs while being constrained by short timelines to bring vaccines to growing markets fast. Under these conditions, a lack of focus on risk-based bioengineering strategies already early in process development will hamper a reliable process design. Therefore, we suggest a risk-based process development framework for bacterial vaccines, incorporating systems metabolic engineering techniques (table 1). Implementing this workflow will require efforts from industry and academia and we propose a synergistic project approach to facilitate a smooth transfer from fundamental finding to actual application in an industrial production process.

Table 1: Applications and limitations of systems metabolic engineering techniques for optimization of bacterial vaccine production processes

Project stage	Technique	Applications	Limitations
Feasibility (design)	Genome-scale model	Comparison metabolic landscapes, rational decision production strain, growth medium development, improvement of yield and robustness	Automatically created models are incomplete, extensive manual curation needed, design defined media not always possible, biomass equation needed
Feasibility (build)	Genome editing / streamlining	Gene knock-ins/knock-outs to improve strain performance, flux re-direction to production pre-cursors virulence factors, improvement of robustness	Physiology of engineered strains hard to predict, requires development genome engineering tools, strict regulations for engineered strains
Feasibility (test)	Transcriptomics	Process control, hypothesis verification	Transcript levels do no always represent protein levels or fluxes
Feasibility (test)	Metabolomics + 13C flux analysis	Model validation, process optimization	No complete metabolome coverage, analysis and sampling techniques are complex
Pre-development (implement/ control)	Genome-scale model	Model predicted control	Steady-state system needed
Pre-development (implement/ control)	Transcriptomics	Process analytical technology, check process consistency	Online analysis methods need to be developed
Pre-development (implement/ control)	Metabolomics	Process analytical technology, check process consistency	For complex/intracellular metabolites online analysis is not yet possible

We consider a vaccine development project where proof-of-concept is developed in academia and will discuss the feasibility phase, where especially genome-scale models can play a role and the pre-development phase, where process controls need to be set based on model predictions and high-throughput (omics) data. We will not discuss the development and registration/commercialization phase because the process must be reliable before the start of the final phases.

Feasibility – Risk-based process development until proof-of-concept

A new vaccine developed in industry is designed according to pre-defined specifications, which are sometimes lacking or not properly described in academic projects[80]. This knowledge gap can seriously delay the transfer of strains developed by bioengineering techniques in academic laboratories to industrial partners[66,81,82] impeding further development and commercialization. Projects in industry follow a risk-based approach where steps are largely standardized using Design For Six Sigma (DFSS[83]) methodology. The most important tool that must be applied to assess risks routinely in each step is failure mode and effect analysis (FMEA[84]). FMEA is a systematic method to analyze where and how a process or design might fail and to assess what the impact of the failure would be in order to pro-actively prevent these risks. In academia, bioengineering projects follow the Design-Build-Test-(Learn) (DBT) scheme[85–87]. DBT and DFSS could be integrated but to be successful there must be early understanding of process risks, assessed by both industrial partners and researchers in academia. This can be achieved by performing a collaborative FMEA before the start of the project, under protection of legal contracts[88], and the risks identified should be captured in the process specifications. After setting the pre-defined process specifications, the initial capabilities of the process need to be determined using a robust experimental design under conditions relevant for the final manufacturing scale. To get this robust experimental design, a flowchart of the anticipated process is needed and process boundaries for each step in the process must be determined using again FMEA. Genome-scale models are essential in this stage of the project to maximize process capability.

Use of genome-scale metabolic models to maximize process capability

Model-based interpretation of biological data is the basis for bioengineering projects and starts with the creation of constraint based genome-scale metabolic models (GEMs). In a bottom-up approach genome information is translated into metabolic capabilities to reconstruct a complete metabolic map of the cell, capturing the stoichiometries, directionalities and gene-protein relationships for all known metabolic reactions and provide a platform for hypothesis-driven investigations, interpretation of multi-omics data and rational strain design[64]. Draft GEMs can be created automatically at low costs based on

the genome sequence of a bacterial strain[62], allowing researchers to start a vaccine development project with a database that links genotype with metabolic capabilities of candidate production strains[89]. GEMs thus represent a species-specific knowledge base. After comparison of metabolic landscapes this knowledge base can be used with advantage to select a production strain, focusing either on reactions important for growth or reactions that play a role in expression of antigens that are often specific virulence factors. This latter aspect requires specific attention as we are just beginning to understand how metabolic networks affect production of virulence factors in pathogenic bacteria[90–94]. When more physiological information is added, the knowledge base will be expanded resulting in a better validated, more predictive metabolic model which allows process developers to predict metabolic fluxes under steady-state approximation and can be applied to rationally design strains that overproduce a target metabolic compound[95]. When this is accomplished, mapping of high-throughput data to the validated model structure provides process developers with a solid basis to understand the process and to analyze consistency in different stages of the vaccine development project. GEMs provide process developers with the knowledge database needed to design experiments to prevent or mitigate process risks identified in further stages of the vaccine development project.

DFSS meets DBT: Rational strain design using GEMs

The Design-Build-Test-Learn cycle applied in academia can be integrated in the Design For Six Sigma (DFSS) methodology (figure 3). Specific integration points are: i) application of GEMs in the measure and analyze phase of DFSS to determine capability and set parameters for the design space, ii) application of the DBT cycle in the design phase of DFSS and iii) model-based interpretation of omics data to generate the final process controls needed during the verify phase of DFSS. Not all bacterial pathogenic strains are considered to be genetically amendable but recent advances in gene editing techniques have expanded the toolbox for researchers to make targeted gene deletions even in bacteria that were previously inaccesible[96,97]. When genetic engineering technologies are applied for improvement of vaccine production strains the rules and regulations, as well as the public acceptance, of the use of these techniques should be considered. In Europe the use of genetically modified organisms (GMO's) is tightly regulated and for each step in the development process from initial production of a strain in the laboratory to application in a clinical trial or field trial, permission should be granted by local authorities and information regarding the safety of the strain used should be available. Changes made to the bacterial genomes should be well documented and checked as even advanced genome-editing techniques are prone to error[98,99].

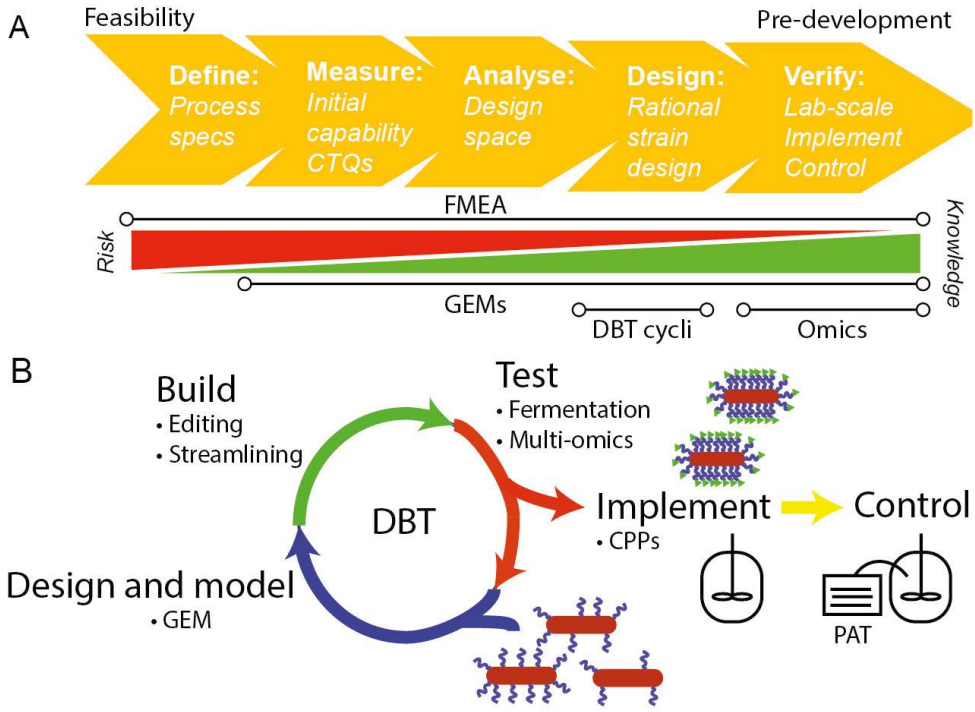


Figure 3. **Design Build Test (DBT) integrated in the Design For Six Sigma (DFSS) methodology.** A: Steps in Design For Six Sigma aligned with bioengineering strategies that should be applied to lower risks and increase process understanding. In the measure and analyse phase genome-scale metabolic models (GEMs) need to be applied to determine process capability, parameters critical to quality (CTQs) and the process design space. Rational strain design should be done following the DBT cycle and in the final stages of the design project omics technologies should be used to develop better process controls. B: The DBT cycle needed for rational strain design. Bioengineering tools that must be applied during each step in the cycle are indicated. The final strain will be used in a vaccine production process after defining critical process parameters (CPPs) and process controls.

Rational strain design for whole-cell inactivated vaccines with undefined antigens

When it is not known which specific antigens are needed, improvement of strain robustness could be achieved by knock-out of non-essential genes predicted by the GEM and verified with gene essentiality screens using random mutagenesis, such as recently performed for the bacterial pathogens *M. pneumoniae*[100], *P.aeruginosa*[101] and *S. pneumoniae*[102]. Alternatively, bottlenecks in metabolism could be removed by overexpression or replacement of key metabolic enzymes or transporters[85]. For non-defined antigens there is a risk that the potency of biomass is diminished after optimizing strains, so strain performance of thus modified strains must be verified under final industrial conditions as well as in an efficacy study.

Rational strain design for bacterial vaccines with defined antigens

When antigens are defined (specific virulence factors on inactivated whole-cells, toxoids or polysaccharide conjugates), a metabolic model must be applied to understand the relationship between metabolic flux and the production of antigenic mass. For increased productivity, process engineers may attempt to increase metabolic flux toward the formation of (precursors for) virulence factors. Again, random transposon mutants libraries will help to assess, with high throughput detection methods (e.g. Tn-seq or HITS)[103], whether genes are needed for growth, antigenic mass formation or both[90]. When a coupling between growth and antigenic mass formation is achieved, algorithms such as Optknock[95] can be used to suggest genetic manipulations to rationally engineer strains with increased antigenic mass production. In addition to these targeted approaches, similar to undefined antigens, genomes could be optimized to get a more robust phenotype. It is also well established that pathogenic bacteria use alternative metabolic pathways while growing in the host and mimicking these growth conditions could result in expression of virulence factors better resembling expression during infection[101,104] and possibly increase yield in fermentor systems.

Rational strain design for live-attenuated vaccines

By simulating a range of growth conditions, GEMs can be applied to predict genes that are either essential or conditionally essential[59] this would allow researchers to make predictions of metabolic reactions that are essential for growth in a host but non-essential for growth in the production process. Disruption of such genes could result in live-attenuated strains with a robust phenotype but a key challenge would be to achieve a balance between attenuation and immunogenicity[105]. Metabolic models cannot be applied to assess immunogenicity and are therefore not suited to predict optimized live-attenuated vaccine strains. When a suitable live-attenuated strain is obtained via alternative methodologies, the genome-scale metabolic model of this strain can be used to rationally engineer more robust strains or strains that reach higher live titers in culture.

Pre-development – Risk-based process control

When a robust production strain is available, a well-designed process control strategy is needed. To determine this control strategy, critical process parameters (CPP's)[106,107] must be defined using a risk-based approach (FMEA) and their influence on product quality must be known. Classical CPP's are the temperature, pH and dissolved oxygen that should be maintained within defined ranges to guarantee product quality. More advanced strain specific CPP's may result from combining high-throughput data generated using multiple omics techniques with knowledge from genome-scale modeling. Insight into novel CPP's could be used to design process analytical tools to better control the process and improve

process reliability. Recent studies showed that especially the combination of metabolomics and transcriptomics profiling of fermentor cultures can be applied to find novel CPP's for bacterial vaccine production. Microarray based transcriptome analysis was successfully applied to find the optimal harvest point, based on expression of virulence factors, during production of antigen for an inactivated *Bordetella pertussis* vaccine[108]. Translation of the transcriptional landscape at maximal antigenic mass levels led to the finding that glutamate and lactate concentrations should not be completely depleted during cultivation which provided a novel CPP. Recently, toxin production in *Clostridium tetanus* was shown to be induced by a metabolic switch from free amino acid consumption to consumption of peptides from complex growth media[109], providing leads for improvement of production media but also enabling additional process control by measuring online concentrations of amino acids or peptides. Further insight into the regulation of expression of virulence factors by constraint-based metabolic models will lead to novel gene clusters to study using a similar methodology. The rapid advances in the field of RNA sequencing and online metabolite analysis would in the near future allow a process engineer to continuously monitor gene expression and metabolite levels in bacterial cultures. Feeding this information into a control model would allow on the spot interventions in the process to optimize expression of antigens. Single-cell RNA sequencing could also elucidate the populations dynamics that exist in a controlled fermentor system and allow local changes to be made to the process to optimize performance of the complete culture. Together with data on the quality of the medium components and process conditions, multivariate data analysis can be used to correlate process data with biomass or antigenic mass yields[110,111].

Concluding remarks

The implementation of systems metabolic engineering strategies in all phases of DFSS process design can mitigate process risk and has a real potential to improve production of complex antigens using pathogenic bacterial strains. But what are the technologies to pick and when should these be applied in the vaccine development project? We have argued that, although antigens are complex and strains and media are not well characterized, there is great potential for applying genome-scale models during process development. These models provide the basis for analysis of metabolic capabilities, understanding flux balances and development of process analytical tools. Bottom-up initiatives for strain design and top-down analysis of high-throughput data could be used to improve process robustness using novel genetic engineering and process analytical tools. However, we like to stress that genetic engineering attempts should be well documented and verified and that industrially relevant conditions should be used to verify process improvements to guarantee a smooth process transfer from academia to industry. We suggest that industry and academia collaborate early in the project, under consultancy agreements, to assess risks in the

production processes and make sure that relevant conditions are used. We strongly encourage further investigation between the coupling of metabolic fluxes and production of virulence factors as this plays a key role in the future systems metabolic engineering framework.

Acknowledgments

We like to thank MSD-AH's global marketing department for providing input regarding the most important bacterial antigens used in animal vaccines.

Chapter 3

Metabolic modeling of energy balances in *Mycoplasma hyopneumoniae* shows that pyruvate addition increases growth rate

Tjerko Kamminga, Simen-Jan Slagman, Jetta J.E. Bijlsma, Vitor A.P. Martins dos Santos, Maria Suarez-Diez and Peter J. Schaap

Published in Biotechnology and Bioengineering

Abstract

Mycoplasma hyopneumoniae is cultured on large-scale to produce antigen for inactivated whole-cell vaccines against respiratory disease in pigs. However, the fastidious nutrient requirements of this minimal bacterium and the low growth rate make it challenging to reach sufficient biomass yield for antigen production. In this study, we sequenced the genome of *M. hyopneumoniae* strain 11 and constructed a high quality constraint-based genome-scale metabolic model of 284 chemical reactions and 298 metabolites. We validated the model with time-series data of duplicate fermentation cultures to aim for an integrated model describing the dynamic profiles measured in fermentations. The model predicted that 84% of cellular energy in a standard *M. hyopneumoniae* cultivation was used for non-growth associated maintenance and only 16% of cellular energy was used for growth and growth associated maintenance. Following a cycle of model-driven experimentation in dedicated fermentation experiments, we were able to increase the fraction of cellular energy used for growth through pyruvate addition to the medium. This increase in turn led to an increase in growth rate and a 2.3 times increase in the total biomass concentration reached after 3-4 days of fermentation, enhancing the productivity of the overall process. The model presented provides a solid basis to understand and further improve *M. hyopneumoniae* fermentation processes.

Introduction

M. hyopneumoniae causes enzootic pneumoniae in pigs. Multiple vaccines are available that provide protection against *M. hyopneumoniae* infection; all contain the inactivated whole-cell bacterium as active component. Manufacturing of these vaccines is done in large-scale fermenter systems in which a sufficiently high biomass concentration should be reached to meet production requirements. However, reaching sufficiently high biomass concentrations in *M. hyopneumoniae* fermentations is challenging due to the fastidious growth requirements of this organism, which remain largely unknown.

Only for a small number of mycoplasma species a chemically defined medium enabling growth is available[75,112,113]. As a result, growth media for production of mycoplasma vaccines are undefined and often contain components from animal origin for which the exact chemical composition is not known such as serum or animal-derived peptones. Moreover, the concentration of critical components such as glucose, lipids and vitamins, varies largely thus challenging the development of a robust production process. In addition, the intrinsic growth rates of mycoplasma species are often low and possibly related to limitations in protein biosynthesis capacity[75], which poses an additional hurdle for the production process.

A constraint-based metabolic model (CBM) provides a list of biochemical reactions, describes the network topology of metabolic pathways and provides a modeling framework to predict and understand biological processes[64]. *Haemophilus influenzae*[114] was the first bacterium for which a CBM was made and since then >400 genome-scale metabolic reconstructions have been made available for the scientific community via the Biomodels database[115]. Considerable efforts have been made to model various mycoplasma species: *M. genitalium*[69], *M. gallisepticum*[68], *M. pneumoniae*[67] and *M. hyopneumoniae* and other swine pathogens[116]. These models show in general that the inferred metabolic networks are small compared to other bacteria, linear and have a low redundancy[75], which is expected in bacteria with a minimal genome.

CBMs are routinely applied to optimize flux through a chosen reaction and to analyze flux distributions that support this minimal or maximal flux given the measured constraints related to substrate uptake or by-product formation. Most often, the selected reaction is biomass formation as flux through the biomass reaction represents the organism's growth rate. Accurate growth rate predictions require detailed knowledge of the chemical composition of biomass. The biomass composition has been determined for *M. pneumoniae*[67], which can be grown on defined medium, but it is a challenge to determine the biomass composition for species that cannot be cultured on defined media.

In this study we developed and deployed a CBM to explore, specifically, energy metabolism. Our goal was to understand and improve a fermentation process used for commercial production of a *M. hyopneumoniae* vaccine. We validated the model with dedicated fermentor experiments in which we measured metabolite profiles and determined

gene expression using RNA sequencing. We used the model to improve aerobic fermentation cultures on complex FRIIS medium thereby showing the potential of our modeling framework for further optimization of *M. hyopneumoniae* vaccine production processes.

Materials and methods

Strain cultivation and genome and transcriptome sequencing

M. hyopneumoniae strain 11 cultures were grown using FRIIS medium[117] to which 1.5 g/l glucose was added or in medium to which 1.5 g/l glucose and 2.2 g/l sodium pyruvate was added. Sartorius Stedim Biostat fermentor systems were used with a maximum capacity of 2.0 L. The pH set-point was controlled at 7.4 using only caustic soda (NaOH, 4N), temperature was controlled at 37°C and the dissolved oxygen concentration was controlled at 5% oxygen saturation using a stirrer speed cascade. Fermentation runs took on average 3-4 days during which 11-14 samples were withdrawn for metabolite and biomass analysis. A flow cytometer (FACSMicroCount, BD) was used to determine the total cell count and based on the titer the biomass concentration was calculated assuming an average cell mass of 0.074 pg/cell[118]. For genome sequencing, DNA was extracted from a bacterial pellet using the Gentra Puregene bacterial kit (Qiagen GmbH, Hilden, Germany). A standardized method (supplementary materials) was used to sequence and assemble the genome using Illumina HiSeq sequencing (paired-end, 50 cycles, 500 mb, 50 bp read length) and PacBio sequencing (1 SMRT cell, 60 mb). RNA was extracted from bacterial pellets using the Qiagen RNeasy mini kit. RNA was sequenced using Illumina HiSeq (single-end reads, half a lane per sample, 50 cycles) using a standardized method (supplementary materials). Gene expression was analyzed using the R Bioconductor package SCAN_UPC[119]. Changes in expression levels per gene were estimated by comparison of RPKM values (Reads Per Kilobase per Million mapped reads). Raw data of genome sequencing and RNA sequencing was deposited in the NCBI Short Read Archive (SRP101540 and SRP053697).

Metabolite analysis using HPLC

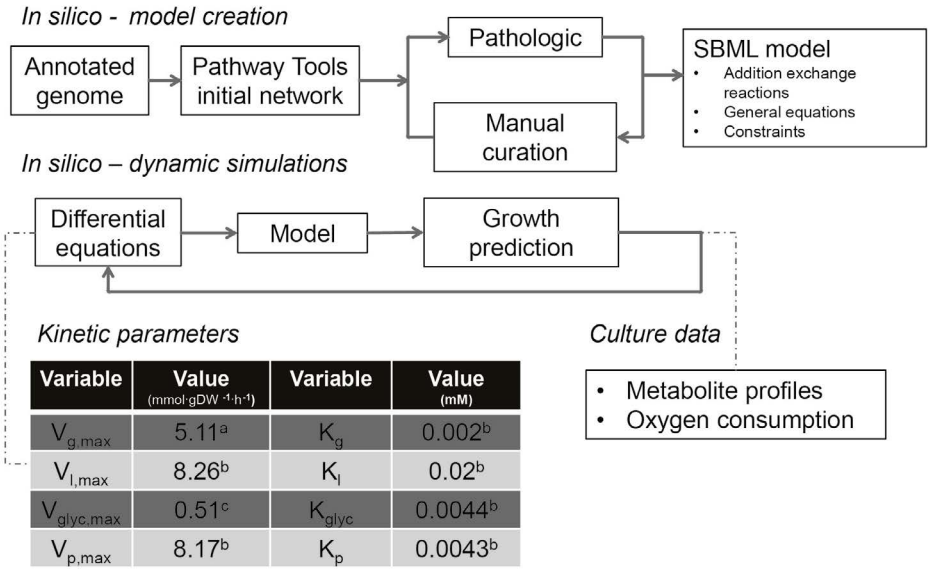
Glycerol, glucose, fructose, mannose, myo-inositol and ribose were quantified in culture supernatant using an ion chromatograph ICS-3000 system with a Dionex CarboPac MA1 (250x4.0 mm) column. Injection volume was 5 µl; eluents used were 100 mM NaOH and 5 mM NaOH. Lactate, acetate, formate and pyruvate were determined using the same HPLC system but fitted with a Dionex IonPac AS-11-HC column (250x4.0 mm) with a 5 µl injection volume and three eluents: 100 mM NaOH, 5 mM NaOH and purified water. Each fermentation sample was measured once per method.

Genome annotation

The genome was annotated using the SAPP annotation pipeline[59]. For gene prediction Prodigal version 2.6.2[120] with codon table 4 was used. Protein annotation was done using InterProScan version 5.17-56.0[58] as described in chapter 4.

Construction genome-scale metabolic model

The annotated genome sequence was imported into Pathway tools[63]. The PathoLogic tool v. 17.0 was used to create a draft metabolic map (figure 1 and supplementary methods), specifically the modules to assign possible enzymatic reactions, protein complexes and to predict presence or absence of metabolic pathways[121]. Manual curation (supplementary materials) was performed using a decision scheme based on the presence or absence of functional protein domains and comparison to annotated genes in the *M. hyopneumoniae* reference genomes (strains: 232, J, 168, 168L, 7422 and 7448) and to the gene content of the *M. pneumoniae* metabolic model[67]. In line with previous studies[75,122], we assumed broad substrate specificity for a number of enzymes in nucleotide metabolism. For instance, pyruvate kinase and the 5'-nucleotidase were assumed to function with 8 nucleotide diphosphates as substrate. The entire Pathway Tools database was exported to flat files and with a BASH script converted to SBML. Cobrapy v0.2.1[123] was used to create a SBML version compatible with MATLAB v.R2016b and the Cobra Toolbox v.2.0.6[124] with LibSBML v.5.13.0 package enabled[125]. Molecular formulas were added for macro-molecules and cellular components when needed (supplementary materials). Additionally, transport reactions were added for all amino acids, nucleotides, vitamins, fatty acids, phosphatidylcholine and general metabolic inputs and by-products for which no annotated transporter existed.



Variable	Value (mmol·gDW $^{-1}$·h^{-1})	Variable	Value (mM)
$V_{g,max}$	5.11[a]	K_g	0.002[b]
$V_{l,max}$	8.26[b]	K_l	0.02[b]
$V_{glyc,max}$	0.51[c]	K_{glyc}	0.0044[b]
$V_{p,max}$	8.17[b]	K_p	0.0043[b]

Fig . 1. Creation and application of a dynamic flux balance model. Steps needed to create the dynamic model from the annotated genome of *M. hyopneumoniae*. Kinetic parameters were obtained from *M. pneumoniae*[67] (a), *M. mycoides*[126] (b) or fitted from *M. hyopneumoniae* HPLC data (c).

Constraint-based modeling

Reactions in the model were represented as a matrix describing reaction substrates, products and their stoichiometries[60]. We assumed biomass consisted of: 62% protein, 20% lipids, 5% DNA, 6.5% RNA, 1.5% free amino acids and 0.1% acyl carrier protein (cofactors and vitamins were not included)[67,127]. Bounds were set for the reaction rates to represent reaction directionality. For several reactions (glucose, glycerol and glycerol-3-phosphate uptake, table S1), the bounds were set to match physiologically known constraints. Protein and RNA degradation were simulated by imposing positive lower bounds to the protein and RNA degradation reactions; values were taken from the *M. pneumoniae* model[67]. Flux balance analysis (FBA) was used to maximize the flux towards the chosen objective reaction, assuming a steady-state system which mimics exponential growth and to identify combinations of fluxes supporting this maximal flux using the Cobra Toolbox v.2.0.6 and Gurobi solver v.6.5.2 (Gurobi Optimization, Houston, TX, USA). Gene essentiality was assessed by setting the reaction rate(s) coupled to a gene of interest to zero and analyzing if the model still supported growth; that is if the model supported flux through the biomass synthesis reaction. Alternative flux distributions were identified through flux variability analysis[128]. Briefly, the maximum value for the objective function was calculated; a new constraint was set for the objective function, so that the flux remains higher than 99.99% (or 99.9% in the second run) of the original value; subsequently each reaction in the model was set as objective function of two FBA rounds maximizing and

minimizing the corresponding reaction. The model was deposited in the BioModels database[129] and assigned the identifier MODEL1704250001.

Dynamic model

Differential equations (supplementary materials) were used to model glucose, lactate, glycerol, oxygen and pyruvate transport using first-order Michaelis-Menten kinetics[130]. Kinetic parameters were obtained from *M. mycoides*[126], *M. pneumoniae*[67] or from batch data of *M. hyopneumoniae* fermentations (figure 1). Time profiles for bacterial cultivations were simulated by combining the set of dynamic equations and the genome-scale metabolic model. The initial substrate, biomass and by-product concentrations were used to calculate the growth rate, substrate and by-product consumption/production rates at t=0 h. With the set of differential equations the changes in substrate and by-product concentrations were calculated using the ode45 solver in MATLAB v.R2016b. In each iterative cycle, the genome-scale model was used to re-calculate the growth rate and substrate and by-product consumption/production rates[61]. Oxygen concentration in the liquid phase was fitted using a second order polynomial equation derived from the experimental data (figure S1).

Results

M. hyopneumoniae strain 11 genome sequencing and comparison to published genomes

Genome assembly including scaffolding and gap-filling resulted in a genome sequence of length 898,877 base-pairs (bp), consisting of 5 scaffolds, of lengths 341707 bp, 244821 bp, 176273 bp, 32027 bp and 4049 bp, with an average GC content of 28.6% (NCBI accession number MWWN00000000). We identified 681 protein coding sequences, 466 proteins with at least one protein domain, 855 unique protein domains and 77 EC numbers using InterProScan (table I). The genome of strain 11 was compared to the currently published *M. hyopneumoniae* genomes (strains: 232, J, 168L, 168, 7422 and 7448) based on the protein domain repertoire[131] as described in chapter 4. We identified a pan-domainome size of 866 protein domains, of which 846 (98%) were present in all strains (core domainome). We conclude from this comparative analysis that the metabolic capabilities of *M. hyopneumoniae* strains were predicted to be highly similar, as was also found by Ferrarini et al.[116].

Table I. *M. hyopneumoniae* genome characteristics.

Strain	Genome size (bp)	Total amount of proteins[a]	Proteins with domains[b]	Total unique domains[c]	Total unique EC#'s[d]
11	898877	681	466	855	77
232	892758	681	468	856	78
J	897405	692	479	859	78
168	925576	698	470	857	78
168L	921093	700	471	857	78
7422	898495	703	477	857	78
7448	920079	690	476	862	78

[a]Amount of proteins based on prodigal gene calling

[b]Number of proteins which contain at least one protein domain

[c]Number of unique domains annotated by InterProScan

[d]Number of unique EC numbers annotated by InterProScan

Genome-scale metabolic model

Based on the genome sequence of *M. hyopneumoniae* strain 11, we created a CBM which consisted of 284 reactions and 298 metabolites (table II, figure 2). The model was termed TK284-MHyo11. On average, the number of reactions catalyzed per enzyme was 1.39 ± 1.05; this value is close to other bacteria with a low number of protein coding genes[75]. Analysis of RNA seq. data using SCAN_UPC[119,132] showed that the genes encoding enzymes underlying the reactions in the model were expressed under aerobic conditions on FRIIS medium (figure S2).

Table II. Model characteristics.

	Model		
Model characteristic	**TK284-MHyo11**	***M. hyopneumoniae* 232**	***M. pneumoniae* M129**
Total reactions	284*	426	306
Number of genes in the model	133	170	145
Number of gene-protein reactions (GPR's)[a]	185	233	197
Orphan reactions[b]	82*	97	92
Exchange reactions[c]	52	97	57
Transport reactions	53	111	75
Metabolites	298	397	266

*Excluded were 9 conversion reactions from mol to gram

[a]Number of gene protein reactions = number of reactions with an associated gene

[b]Orphan reactions = reactions without an associated gene, not including exchange reactions

[c]Exchange reactions = reactions needed to exchange consumed or secreted compounds

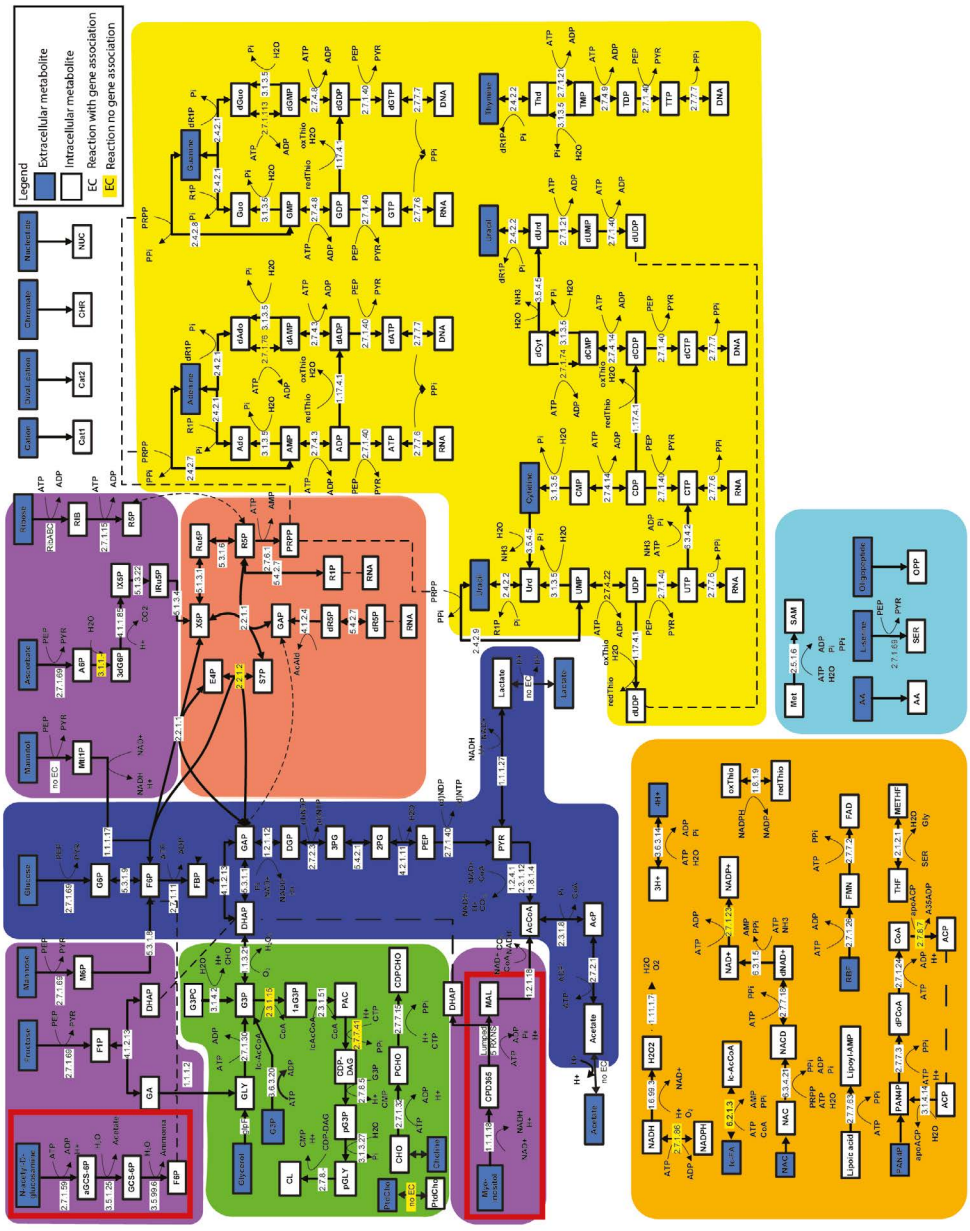


Fig. 2. Metabolic map of *M. hyopneumoniae*. Pathways are indicated by background color: glycolysis (blue), alternative sugar metabolism (purple), lipid metabolism (green), vitamin and co-factor metabolism (orange), nucleotide metabolism (yellow), PPP (red) and amino acid metabolism (blue). Pathways highlighted in red are absent from *M. pneumoniae*. Abbreviations are explained in table S11.

Comparison to existing mycoplasma models

The number of reactions in our model was lower than the number found in the model of *M. pneumoniae* or the reference model of *M. hyopneumoniae* (encompassing 306 and 426 reactions respectively, table II). The functional annotation of our model on the other hand was compared to the reference model of *M. hyopneumoniae* strain 232 and was found to be better aligned with the genome information as it contained less (orphan) reactions unlinked to genome encoded functions. This was a direct result of the approach used during manual curation, where we assigned functionality based on the presence of protein domains associated to a metabolic reaction and, in general, only added reactions for which associated protein domains where found in the genome, which resulted in a reduction in the number of orphan reactions. For example, we did not find protein domains related to fumarate production, succinate production, folate metabolism, phosphatidylcholine synthesis and synthesis of glycolipids and therefore did not incorporate these reactions in our model although they were present in the reference model as orphan reactions. Interestingly, in the reference model of *M. hyopneumoniae* strain 232 there was no annotated gene for alcohol dehydrogenase (EC 1.1.1.1), while other *M. hyopneumoniae* strains did have this functionality annotated. In strain 11 we did not find a gene for alcohol dehydrogenase either. Comparison of our *M. hyopneumoniae* metabolic model to the *M. pneumoniae* metabolic model showed notable differences such as the presence (in our model) of a myo-inositol degradation pathway, the presence of a N-acetyl-D-glucosamine degradation pathway and a likely absence of folate metabolism as only a single gene related to folate metabolism could be found in the strain 11 genome: EC 2.1.2.1: glycine hydroxymethyltransferase. Also absent in our model when compared to *M. pneumoniae* were alcohol dehydrogenase, arginine fermentation, glycolipid metabolism and thymidylate synthase.

Parameterization and validation

We measured the initial substrate concentrations in several medium lots to determine starting concentrations of metabolites (table S2). The average growth rate reached in FRIIS medium was 0.0177 h^{-1} (duplicate batches 0.0219 h^{-1} and 0.0135 h^{-1}, respectively). Part of the growth rate variation between batches is most likely caused by variations in the medium composition due to the use of components from animal origin. For initial model simulations the only carbon sources for which consumption was allowed were glucose, glycerol and glycerol-3-phosphate. We had no analytical method to measure glycerol-3-phosphate in the medium and therefore copied the consumption rate for this compound from the *M. pneumoniae* model. Additionally, we allowed *in-silico* consumption of other components such as amino acids, nucleotides and lipids as listed in table S3. We used the CBM to estimate (and subsequently fix in the model) an energy expenditure of 18.4 $mmol \cdot gDW^{-1} \cdot h^{-1}$ as non-growth associated maintenance rate (NGAM). The NGAM was estimated by adding a non-specific energy consumption reaction to the model. The lower limit of this reaction

was then gradually increased until agreement was found between the measured and the model predicted growth rate. The parameterized CBM was used to build the integrated dynamic model, which was validated by simulation of time profiles in aerobic fermentor cultures. We assumed glycerol to be a non-limiting component in these cultures since it could be obtained from complex medium components. Based on oxygen consumption profiles we calculated an average lag-phase of 12.5 hours for the duplicate fermentation cultures and started model simulations with a biomass concentration of $5.89 \cdot 10^{-3}$ $gDW \cdot l^{-1}$. Measured biomass and glucose concentrations were in agreement with predicted values (figure 3A and 3B). The predicted acetate concentration was slightly lower than measured (figure 3C), indicating possible consumption of additional carbon sources from the complex medium.

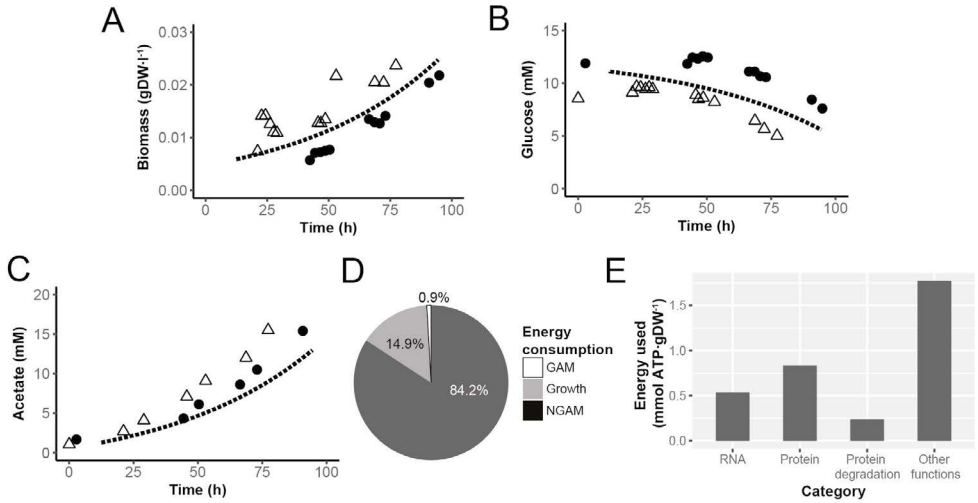


Fig. 3. Metabolite profiles and energy balances in standard *M. hyopneumoniae* fermentations. Comparison of model predicted aerobic batch profiles (dashed line) to measured profiles in fermentation batches (N=2, dots and pyramids) for: A: biomass; B: glucose; and C: acetate concentration. D: distribution of energy consumption. E: major growth related energy consumption.

Analysis of in silico flux distributions and energy balances

Up to 124 enzymatic and transport reactions were able to carry flux while growing on glucose and glycerol under aerobic conditions, 44% of the total amount of reactions in the model (excluding conversion reactions, table II). NGAM required 84.2% of total cellular energy, 0.9% was used for growth-associated maintenance and 14.9% for reactions needed for growth (figure 3D, table S4). In *M. pneumoniae* cultures 71-88% of cellular energy was used for NGAM and 12-29% for growth, depending on the culture stage[67]. Among the growth related functions, the ones consuming the most energy were protein production, RNA production and protein degradation (figure 3E). Energetic costs of DNA production

were negligible compared to these other categories. When compared to the growth related energy sinks in *M. pneumoniae*, our model did not contain protein folding as energy sink and had reduced lipid synthesis capability.

Flux variability analysis and gene essentiality analysis

We performed flux variability analysis to explore alternative flux distributions compatible with at least 99.99% of the maximal growth rate. The majority of the reactions showed negligible variability as was also found for *M. pneumoniae*[67]. We identified the 20 reactions that showed the largest flux change (table S5). All these reactions were related to conversion of purine nucleotides; there was some flexibility in the network which allowed for alternative pathways to produce dADP and dGDP (figure 2). Reaction directionality was changed for the reactions catalyzed by phosphoglycerate kinase which confirmed that the flexible substrate use assumed for this enzyme improved network flexibility. We also performed flux variability analysis allowing 0.1% (table S6) variation in the growth rate and we identified additional reactions for which directionality changed: CMP kinase and the asparagine-tRNA ligase. For the latter function there were two separate reactions in the model, one irreversible and one reversible reaction which provided flexibility in the network and allowed the direction change for the reversible reaction. Finally, gene essentiality analysis showed that 41% of genes in the model were classified as essential for growth (table S7). There is currently insufficient experimental information on gene essentiality in *M. hyopneumoniae* to validate this result[133].

Model predicts that pyruvate addition increases the growth rate

As was also observed in *M. pneumoniae,* the metabolic pathway that carried the largest amount of flux was glycolysis and pyruvate metabolism. Pyruvate has been mentioned in literature to increase the growth rate of mycoplasma species[134]. We confirmed that in our model pyruvate addition increases growth rate and ran dedicated fermentor studies to assess the effect on growth and metabolite profiles (figure 4A-E). Interestingly, we observed production of lactate under aerobic conditions (figure 4D) which is energetically unfavorable. Three metabolic scenarios were simulated with the dynamic integrated model: i) growth on pyruvate with lactate production and a lowered biomass specific glucose uptake rate, ii) growth on glucose and pyruvate and iii) growth on lactate, glucose and pyruvate. In the initial metabolic condition the flux through the PEP-PTS transporter for glucose was assumed to be lower as a result of product inhibition. The previously fitted NGAM rate was too high to support growth in the initial metabolic condition when cells were growing on pyruvate and glucose. The growth rate measured in the initial stage of the culture was compatible with a decreased NGAM rate of 10.15 mmol $ATP \cdot gDW^{-1} \cdot h^{-1}$. Based on measured metabolite profiles we increased the flux through the glucose transporter when the pyruvate concentration was less than 2.7 mM. This had a positive influence on the growth rate (figure 4A). Near the time of pyruvate exhaustion in the medium, lactate is

consumed and growth-rate further increases. In our simulation, the culture reached stationary phase when a biomass concentration of 0.05 $gDW \cdot l^{-1}$ was reached according to measured biomass data. Energy balances at early and late (t=45h) exponential growth showed that in the simulations with pyruvate addition, a higher fraction of energy was allocated to growth (figure 4F).

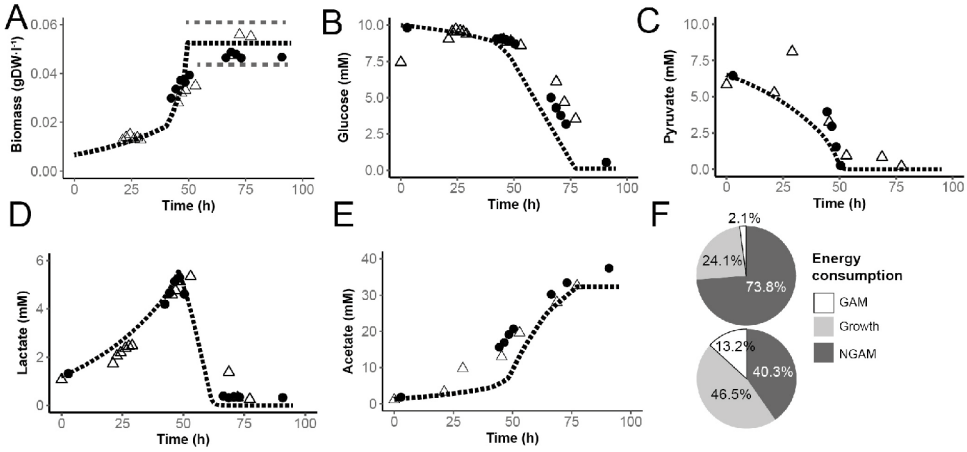


Fig. 4. Pyruvate supplementation increases total biomass yield and results in lactate production under aerobic conditions. Comparison of model predicted aerobic batch profiles (dashed line) to measured profiles in fermentation batches (N=2, dots and pyramids) after pyruvate supplementation to medium for: A: biomass, dotted grey lines indicate variation in the final concentration of biomass reached, B: glucose. C: pyruvate. D: lactate and E: acetate concentration. F: energy distribution early stage (top) and at t = 45h.

Strain design

The model was used to study the impact of gene knock-in mutants on growth assuming growth on FRIIS medium with only glucose added. Our model includes a complete myo-inositol consumption pathway which was absent in other mycoplasma species. We did not observe consumption of this component in aerobic fermentor studies and in our model the myo-inositol transporter is an orphan reaction as there was no annotated transporter. A *M. hyopneumoniae* mutant with a knock-in of a myo-inositol permease should be able to consume myo-inositol from the growth medium, for such a mutant the model predicted a doubling of the growth rate (from 0.0177 h^{-1} to 0.0352 h^{-1}) upon consumption of 0.25 mmol myo-inositol$\cdot gDW^{-1} \cdot h^{-1}$ (table S8). If a knock-in with an ABC transporter for myo-inositol is done, the ATP demands of the transporter would enable a slightly lower growth rate (0.0308 h^{-1}). An even more challenging knock-in would be the expression of the complete pathway for arginine fermentation, which is absent in *M. hyopneumoniae*. In *M. pneumoniae* this pathway consists of five genes (MPN304-307 and MPN560) coupled to three reactions. Simulation of a knock-in with the five *M. pneumoniae* genes and an additional transporter for one of the by-products L-ornithine leads to an increased *in silico*

mutant growth rate of 0.0235 h^{-1} (table S8), as long as arginine consumption is allowed with the same flux as determined in *M. pneumoniae*.

Discussion

We have developed and applied a CBM of *M. hyopneumoniae* strain 11 to study and improve biomass yield in fermenter systems. Our aim was to understand cellular energy balances during growth and find methods to increase biomass yield. We found that in the standard cultivation medium used for *M. hyopneumoniae*, 84% of total ATP production by the cell is used for non-growth associated maintenance and only 14.9% is used for growth. This confirms the results found in *M. pneumoniae*. A possible explanation for the high percentage of energy used for NGAM could be the high surface-to-volume ratio of mycoplasma species which means that maintaining cellular homeostasis is energetically expensive[67].

Because our focus for modeling was mainly on understanding cellular energy balances we used a simplified biomass composition in the model. We calculated ATP consumption of reactions related to vitamin and cofactor metabolism in the *M. pneumoniae* model and found that these reactions consumed only 0.003% of total cellular energy (table S9). This shows that strictly for understanding the cellular energy distribution, the contribution of reactions in co-factor and vitamin production can be neglected. Moreover, it remains unclear which of these components are consumed directly from the complex medium and which are produced by the bacterium. Our CBM also lacked almost all reactions related to folate metabolism although this component is generally assumed to be essential for growth because formylation of initiator methionyl-tRNA is needed to start translation in bacteria[135]. However, because formylation of methionyl-tRNA does not occur in *Pseudomonas aeruginosa*[136] and is also not described as part of the minimal bacterial gene set[137], we find the assumption that it is absent in *M. hyopneumoniae* reasonable.

Our simulations show that a large part of the metabolic network is not used for growth, in the simulated conditions, and as a result many genes are predicted to be non-essential. This is an interesting result as mycoplasma genomes are expected to be close to the minimal gene set required for life without a host and therefore a high percentage of genes were expected to be essential. Part of the result can be explained because we applied a simplified biomass equation, for example, addition of NADPH, FAD and CoA to the biomass equation added 5 essential genes, a 3.8% increase in the percentage of essential genes (table S7). Also our predictions for gene essentiality were based on growth in complex medium while conditions in the host will be different and gene essentiality might be higher because nutrients are scarce. Though not all genes were found to be essential, our transcriptomics dataset indicated that all genes in the model were expressed and although the transcriptome dataset collected is small, we identified a correlation between gene

expression and flux for the glycerol uptake facilitor (glpF) and for the glucose PTS transporter (MHP629) (table S10).

Given the challenges associated with introducing genetic modification in *M. hyopneumoniae*, the potential of the described knock-in mutants has to be carefully evaluated. Our model predicted that both the introduction of a myo-inositol transporter (single gene knock-in) as well as an arginine fermentation pathway (5 gene knock-in) potentially increases the growth rate of mutant strains.

To our knowledge, this is the first study where *M. hyopneumoniae* growth is studied in controlled fermenter systems. In these systems pH and oxygen concentration were controlled resulting in a stabilized growth profile and allowing accurate measurement of growth rate, substrate uptake rates and by-product formation which is essential data for model validation and parameterization. Our stringent approach for model creation, applying annotation based on functional protein domains, resulted in a more condensed model of *M. hyopneumoniae* when compared to the reference model. Following an iterative cycle of model driven experimentation, we were able to double biomass production and reduce total process time by pyruvate addition to aerobic fermentations, which improved the economic potential of the production process. The model we present provides a solid basis to understand the metabolic capability of *M. hyopneumoniae* and further optimize the production process regarding biomass yield and process robustness.

Nomenclature

CBM: Constraint-based model; SBML: Systems Biology Markup Language; COBRA: Constraint Based Reconstruction and Analysis; DW: Dry weight; BASH: Bourne-again shell; Metabolite abbreviations are explained in table S11.

Acknowledgements

Mark Davids for help with model transformation from Pathwaytools to SBML and further adaptation of the model using the COBRA toolbox in Matlab.

Chapter 4

Persistence of Functional Protein Domains in Mycoplasma species and their role in Host Specificity and Synthetic Minimal Life

Tjerko Kamminga, Jasper J. Koehorst, Paul Vermeij, Simen-Jan Slagman, Vitor A.P. Martins dos Santos, Jetta J.E. Bijlsma and Peter J. Schaap

Published in Frontiers in Cellular and Infection Microbiology

Abstract

Mycoplasmas are the smallest self-replicating organisms and obligate parasites of a specific vertebrate host. An in-depth analysis of the functional capabilities of mycoplasma species is fundamental to understand how some of simplest forms of life on Earth succeeded in subverting complex hosts with highly sophisticated immune systems.

In this study we present a genome-scale comparison, focused on identification of functional protein domains, of 80 publically available mycoplasma genomes which were consistently re-annotated using a standardized annotation pipeline embedded in a semantic framework to keep track of the data provenance. We examined the pan- and core-domainome and studied predicted functional capability in relation to host specificity and phylogenetic distance.

We show that the pan- and core-domainome of mycoplasma species is closed. A comparison with the proteome of the "minimal" synthetic bacterium JCVI-Syn3.0 allowed us to classify domains and proteins essential for minimal life. Many of those essential protein domains, essential Domains of Unknown Function (DUFs) and essential hypothetical proteins are not persistent across mycoplasma genomes suggesting that mycoplasma species support alternative domain configurations that bypass their essentiality.

Based on the protein domain composition, we could separate mycoplasma species infecting blood and tissue. For selected genomes of tissue infecting mycoplasmas, we could also predict whether the host is ruminant, pig or human. Functionally closely related mycoplasma species, which have a highly similar protein domain repertoire, but different hosts could not be separated. This study provides a concise overview of the functional capabilities of mycoplasma species, which can be used as a basis to further understand host-pathogen interaction or to design synthetic minimal life.

Introduction

Mycoplasmas have evolved from a common gram-positive ancestor[14] and the evolutionary path of genome reduction has led to an obligatory parasitic lifestyle which presumably has selected for those bacteria that best manipulate their hosts and make optimal use of their specific niche with a minimal set of genes. The mechanisms needed by these bacteria to survive in a vertebrate host, however, are not completely understood[138]. Research into infectious mechanisms used by mycoplasma species has been focused on identification of adhesive molecules[48], lipoproteins[139], molecular mechanisms used to vary the composition of the surface of the bacterial membrane[14] and production of oxidizing components (e.g. hydrogen peroxide and hydrogen sulfide) which cause damage to the host[36,140]. While these studies provide insight into the mechanisms used by mycoplasmas to infect the host they do not explain why a mycoplasma species is specific for its host. Besides being important pathogens, mycoplasma species have also been extensively studied because their gene set is expected to be close to the minimal amount of genes needed to sustain life[137]. Recently, a major hallmark was achieved by publication of an engineered mycoplasma with a synthetic minimal genome of 473 genes based on the genome of *Mycoplasma mycoides subsp. capri*[86,141] providing a benchmark for genome comparison studies aimed at determining gene essentiality.

Advancements in genome sequencing techniques led to the availability of a multitude of genomes from mycoplasma species. With this wealth of sequencing data, it is possible to study the complete repertoire of genes for a bacterial species, the pan-genome. Rouli *et al.*[142] observed that bacterial species that have adopted an allopatric lifestyle in specific hosts, tend to have a closed pan-genome. In recent comparative genomics studies for mycoplasma and haemoplasma species, a sub-group within the mycoplasma genus, the pan-genome was reported to be open[143,144]. Here we present a genome-scale comparison of mycoplasma species at the functional level of protein domains. Proteins are the main working machinery of the cell and consist of functional domains, which are stable structurally independent and genetically mobile units. A protein function can thus be precisely described by taking into account the specific domain composition architecture[131]. Studying protein domain presence, instead of gene sequence similarity, allows for comparison of domain promiscuity and expansion and domain architecture variability. In a recent study, this approach was used for comparison of 121 *Streptococcus* strains based on the protein domain composition of these strains[145] and the authors were able to capture metabolic flexibility within *Streptococcus* through the identification of differences in core metabolic pathways between pathogenic and non-pathogenic strains. By analyzing functional capability based on protein domains, we gain insight in functional flexibility of mycoplasma species and we hypothesized that this will allow us to capture functional differences between mycoplasma species explaining adaptation to a host or niche. This strategy is supported by the recent finding that for *Mycoplasma pneumonia* gene

essentiality should be studied on the level of domains and not on the level of genes[100]. All protein domains found in a species make up the pan-domainome of a species[146], containing core, accessory and unique domains.

We performed a de novo annotation of 80 publically available mycoplasma genomes and included in this analysis the synthetic minimal genome variant of *M. mycoides subsp. capri* using a standardized pipeline for prokaryotic genomes focused on identification of protein domains. We determined the composition and size of the core- and pan-domainome of distinct mycoplasma species and of the complete mycoplasma genus. Incorporation of the synthetic minimal variant in the comparison allowed us to analyze the overlap between protein domains in the core domainome of mycoplasma species versus the synthetic minimal bacterium to pinpoint functions essential for minimal life.

Methods

Genome retrieval and data handling

In total 65 complete and 15 draft mycoplasma genomes (table S1) were obtained from the NCBI database on the 25th of August 2015 using the "rsync" interface. The dataset contained information from 34 mycoplasma species. For 20 species a single genome sequence was available while for the other 14 species multiple genomes were available (2-12 genomes per species). For 6 species only a draft genome sequence was available. Genome sizes range from 0.58 Mbp for *M. genitalium* to 1.36 Mbp for *M. penetrans*. Genome sequences were retrieved in FASTA format and were used as input for an in-house prokaryotic annotation platform (SAPP[59]). *Bacillus subtilis* strain 168 (NC_000964)[147] was used as outlier/common ancestor. Briefly, the SAPP platform consists of sets of modules required for genome annotation of prokaryotes. Different modules can be selected for analysis and results and metadata are directly stored in a graph-database using the RDF (Resource Description Framework) data model. Originally deposited genome annotations were obtained directly from the NCBI in GenBank format and converted into RDF. For three draft genomes no reference annotation was available (accession numbers: NZ_ANIV00000000, NZ_ANAB00000000 and NX_ANAA00000000).

Genome re-annotation using SAPP

Gene prediction was performed using Prodigal version 2.6.2[120] with codon table 4 (The Mold, Protozoan, and Coelenterate Mitochondrial Code and the Mycoplasma/Spiroplasma Code). Proteins were analyzed using InterProScan version 5.4-47.0[58] with the complete set of applications enabled (TIGRFAM, PIRSF, ProDom, SMART, PROSITE Profiles&Pattern, HAMAP, PfamA, PRINTS, SUPERFAMILY, Coils and Gene3D). Protein domain information and other relevant information (GO terms, EC#'s) obtained from InterProScan were directly stored in the graph-database. For querying results a

SPARQL endpoint was set-up on a local server using Blazegraph Workbench v2.1.0. The annotated genomes were uploaded in RDF format using the Blazegraph Webinterface and query results were obtained in R using RCurl[148] and SPARQL[149].

Phylogenetic analysis of mycoplasma genomes

16S rRNA sequences were obtained from the ARB-SILVA database[150] (table S2). When available, sequences from the "all species living tree" project were used. For the synthetic JCVI-Syn3.0, the 16S rRNA sequence of the parental *M. mycoides subsp. capri* PG3 was used. 16S rRNA sequences were aligned with Clustal Omega (version 1.2.1). MEGA (version 7.0.14) was used to create a phylogenetic tree (maximum likelihood method with 500x bootstrapping). Archaeopteryx (version 0.9901) was used to visualize the tree and root the tree using *B. subtilis* as outlier. The phylogenetic tree was read into R and analyzed using the R package "ape"[151]. Comparison of the phylogenetic tree to the protein domain tree was done using the R package "dendextend"[152].

Analysis of core- and pan-domainome

The total domain composition of each genome was obtained using SPARQL queries. Only domains which were assigned with an e-value of $<1E^{-07}$ were taken into account. In R, a matrix was created with all genomes and their domain composition in binary format, meaning that in this analysis only domain presence was considered. Clustering of species based on the presence/absence matrix was done using the function "hclust" in R; distances were calculated using the "Manhattan" distance. The R-package "micropan"[153] was used to analyze the pan- and core-domainome of species from which five or more genomes were available and the same approach was used to analyze the complete mycoplasma database. To analyze how the amount of genomes sequenced affects the size of the pan- and core-domainome, a 10 times random sampling was done from the presence/absence domain matrix using sample sizes ranging from 1 to 80 genome sequences. The range of model complexities considered (k-range) was 3-5. Estimated core- and pan-domainome sizes were obtained using micropan; true core- and pan-domainome sizes were directly calculated from the sample set. Further analysis of differences between species was done using principal component analysis (PCA). Loading scores obtained with PCA were used to identify domains that contribute highly to group separation. To identify domains present in haemoplasma species that contribute highly to separation of this cluster from the other mycoplasma clusters a loadings score >0.02 was used. To identify domains that contribute highly to separation of the pneumoniae cluster and the spiroplasma/hominis cluster, cut-off values for the loading score of >0.05 and <-0.05 were used, respectively. Proteins with a metabolic function were extracted from the genome-scale metabolic model of *M. pneumonia*[67] and extended with InterProScan domain annotations.

Analysis of orthologous proteins

A SPARQL query was used to generate a protein FASTA file using all mycoplasma genomes (JCVI-Syn3.0 was not taken into account). An all-against-all BLASTP[154] was performed of the mycoplasma proteins using an e-value cut-off of $1E^{-05}$ and a maximum target sequence of 10^5. The BLAST file created was used to find orthologous proteins with orthAgogue[155] excluding protein pairs with an overlap below 50%. Clustering of orthologous proteins was done using MCL[156] setting 1.5 as main inflation. With a SPARQL query the domain composition of all orthologous proteins was obtained based on InterPro identifiers.

Clustering of hypothetical mycoplasma proteins

Hypothetical proteins (domain-less proteins) were obtained from the mycoplasma genomes using a SPARQL query. Orthologous protein clusters containing these hypothetical proteins were obtained from the list of orthologous protein clusters. Persistence of these orthologous clusters was determined in the complete set of genomes used, JCVI-Syn3.0 and *M. mycoides subsp. capri LC*. Haemoplasma species were excluded from this analysis.

Prediction of host/niche specific domains

K-nearest neighbor and random forest classification[157] were used to classify mycoplasma species based on host or niche specificity and to identify domains important for classification. A binary domain presence/absence matrix was used as input. The R-package "class" was used to perform k-nearest neighbor (k-nn) classification[158] and the R package "randomForest"[159] was used for random forest classification. 500 trees were built for each classification with random forest. Domains important for classification were found based on the mean decrease in node impurity (Gini index). Information from 26 mycoplasma genomes was used for the final niche classification and from 22 genomes for the final host classification (tables S10 and S11). K-nn classification for the niche dataset was done with a k-value of 5, 19 mycoplasma species in the training set and 7 mycoplasma species in the test set (4 infecting multiple tissue types and 3 infecting strictly respiratory tissue). For the host classification a k-value of 4 was used, 16 mycoplasma species in the training set and 6 mycoplasma species were used in the test set (3 species infecting ruminants, 1 species infecting pig and 2 species infecting humans).

Results

Re-annotation of mycoplasma genomes

The quality of the structural and functional annotation of publically available genomes can vary. In order to get for all studied genomes an up-to-date set of annotated genes and to minimize the risk of false discoveries resulting from methodological inconsistencies, the 80 publicly available mycoplasma genomes (table S1) were re-annotated using a standardized set of algorithms[59]. Mycoplasma genomes have a low GC content and thus accuracies of the various original gene prediction methods applied were expected to be high [160]. Nevertheless, on average 3.7% more genes were found after re-annotation with the most recent version of prodigal[120]. Consequently, the total number of proteins found in the re-annotated genomes was also higher (table S3).

Mycoplasma proteome and (predicted) pan- and core-domainome

Haemoplasma species, which specifically infect blood, have a higher number of predicted proteins relative to their genome size (Fig 1A), corresponding with a lower average CDS length[161]. This difference is caused by the presence of a large repertoire of proteins with a relatively short CDS length, which are part of paralogous gene families[162]. To survive in their specific niche, haemoplasma species can express these proteins and generate variability of proteins at the cell surface to prevent detection by the immune system of the host[72]. Besides the haemoplasma species, a high amount of predicted proteins relative to the genome size was also found in *M. genitalium* G37 and JCVI-Syn3.0. Approximately 80% of the mycoplasma species proteins contained functional domains (Fig. 1B). This percentage is similar to the average match percentage found if the whole UniProtKB is analyzed using InterProScan[163]. The *M. mycoides* based JCVI-syn3.0 synthetic genome contained the highest percentage of proteins with a recognizable domain (86%), approximately 9% more than the parental template genome. Haemoplasma species were the notable exceptions, which despite their normal genome size, contained a significantly lower percentage of proteins with recognizable domains (22-54%). This difference occurs because the aforementioned variable surface proteins do not contain recognizable domains. As a direct result of the re-annotation strategy the total amount of unique functional domains per species increased with 0.8% on average.

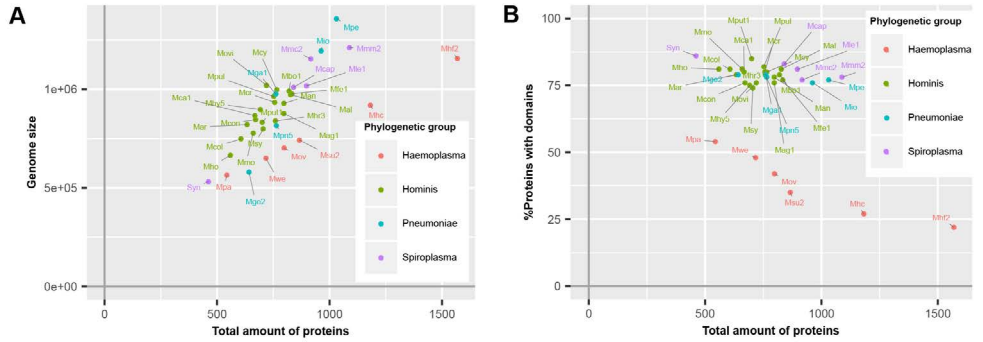


Fig. 1. Mycoplasma proteome specification. A: Correlation between genome size and total amount of proteins. B: Ratio of proteins covered by protein domains. Abbreviations used: Mag, *M. agalactiae*; Mal, *M. alligatoris*; Man, *M. anatis*; Mca, *M. canis*; Mcol, *M. columbinum*; Mge, *M. genitalium*; Mio, *M. iowae*; Mmc, *M. mycoides capri*; Movi, *M. ovipneumoniae*; Mpn, *M. pneumoniae*; Mar, *M. arthritidis*; Mbo, *M. bovis*; Mca, *M. capricolum* subsp. *capricolum*; Mcon, *M. conjunctivae*; Mcr, *M. crocodyli*; Mcy, *M. cynos*; Mfe, *M. fermentans*; Mga, *M. gallisepticum*; Mhc, *M. haemocanis*; Mhf, *M. haemofelis*; Mho, *M. hominis*; Mhy, *M. hyopneumoniae*; Mhr, *M. hyorhinis*; Mle, *M. leachii*; Mmo, *M. mobile*; Mmm, *M. mycoides* subsp. *mycoides*; Mov, *M. ovis*; Mpa, *M. parvum*; Mpe, *M. penetrans*; Mpul, *M. pulmonis*; Mput, *M. putrefaciens*; Msu, *M. suis*; Msy, *M. synoviae*; Mwe, *M. wenyonii*; Syn, JCVI-Syn3.0. Numbers relate to strains (Table S4).

The total pan-domainome consisted of 1737 domains, the core domainome consisted of 335 domains and the core-to-pan ratio was 19.3%. Analysis of the pan-domainome for species from which 5 or more genomes (*M. pneumoniae, M. gallisepticum, M. hyopneumoniae* and *M. genitalium*) were available using Micropan (2- or 3-component system[153]) showed that the pan-domainome was closed (alpha>1). A closed pan-domainome was also observed for the genus (9 component system, alpha>1) taking into account all 80 mycoplasma genomes (figure 2).

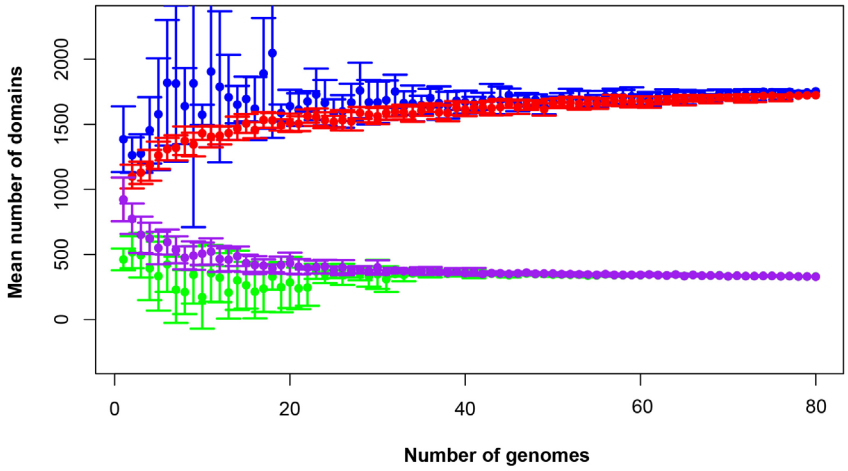


Fig. 2. The mycoplasma pan-domainome is closed. True and estimated core- and pan-domainome were calculated using an iterative process (n=10) in which a fixed number of genomes was randomly selected from the binary domain matrix. Estimated values were calculated using the R package MicroPan and true values were directly calculated from the input. Estimated pan-domainome (blue); true pan-domainome (red); estimated core-domainome (green) and true core-domainome (purple).

Functional classification of mycoplasma species

To gain insight into a possible functional differentiation of mycoplasma species as a result of specific host co-evolution, we clustered mycoplasma species based on a presence/absence domain matrix and compared domain repertoire clustering with clustering based on 16S rRNA sequences (figure 3). In the domain based functional tree, the monophyletic pneumonia cluster separated into three separate functional clusters. One of these separate clusters contains the haemoplasma species, which have a relatively low number of protein domains (figure S1 and table S4). *M. penetrans* and *M. iowae* form a second functional cluster; these species have a relatively high number of functional domains when compared to other species in the pneumonia 16S-phylogenetic group. The remaining species in this 16S-phylogenetic group are closely related to the spiroplasma cluster in the functional tree. The hominis 16S-phylogenetic cluster was completely maintained in the protein domain tree but compared to the 16S tree there were some re-arrangements, which can partly be explained by low significance in the assignment of branches in the 16S phylogenetic tree. Notable changes are: *M. hominis* and *M. arthritidis* clustered with *M. columbinum* and *M. pulmonis* clustered with *M. hyorhinis*. We did not observe a functional clustering based on host.

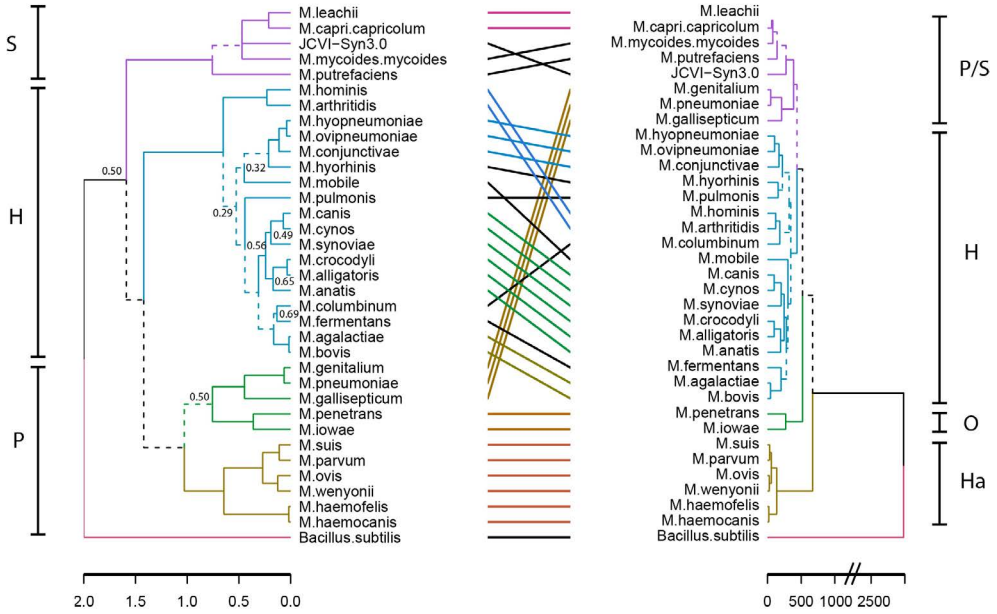


Fig. 3. Niche-driven functional evolution. Accelerated functional evolution causes separation of haemoplasma species and several other mycoplasma species when phylogenetic clusters are compared to functional clusters. Dashed lines indicate distinct branches. Left: standard phylogenetic tree using 16S rRNA (maximum likelihood, 500x bootstrapped, see S2 for strains and sequences which were used). Right: functional clustering based on Manhattan distance calculated from the presence/absence matrix of domains. Groups indicated are: S, Spiroplasma; H, Hominis; P, Pneumoniae; Ha, Haemoplasma and O, Other.

Functional differentiation of haemoplasma species

To determine which domains were important for separation of haemoplasma from mycoplasma species infecting tissue, we used principal component analysis (figure 4). Based on the loading scores for the first and second principal component we could assess which domains contributed to group separation. Haemoplasma species were separated from the other mycoplasma species along the first principal component. We identified 30 domains in haemoplasma species that mainly contributed to separation of this cluster (table 1 and table S5) and 400 domains present in the tissue infecting mycoplasma species that mainly contributed to separation of this cluster from the haemoplasma species cluster. Domains present in the haemoplasma species that contributed to group separation were ABC transporter domains for iron or vitamin B12. Multiple domains were found related to functional enzymes in purine metabolism (GMP synthase, IMP dehydrogenase, adenylosuccinate synthase) or L-aspartate metabolism (fumarate lyase family domains, part of adenylosuccinate lyase) which provides a precursor for purine metabolism[164]. The presence of GMP synthase domains may provide the haemoplasma with the option to

produce all purine bases from hypoxanthine which is present in blood[161]. An alternative function for these GMP synthase domains could be the production of glutamate which is present in a low concentration in blood[165]. Three domains related to superoxide dismutase activity were also found, a function, which could provide protection when radicals are present in blood.

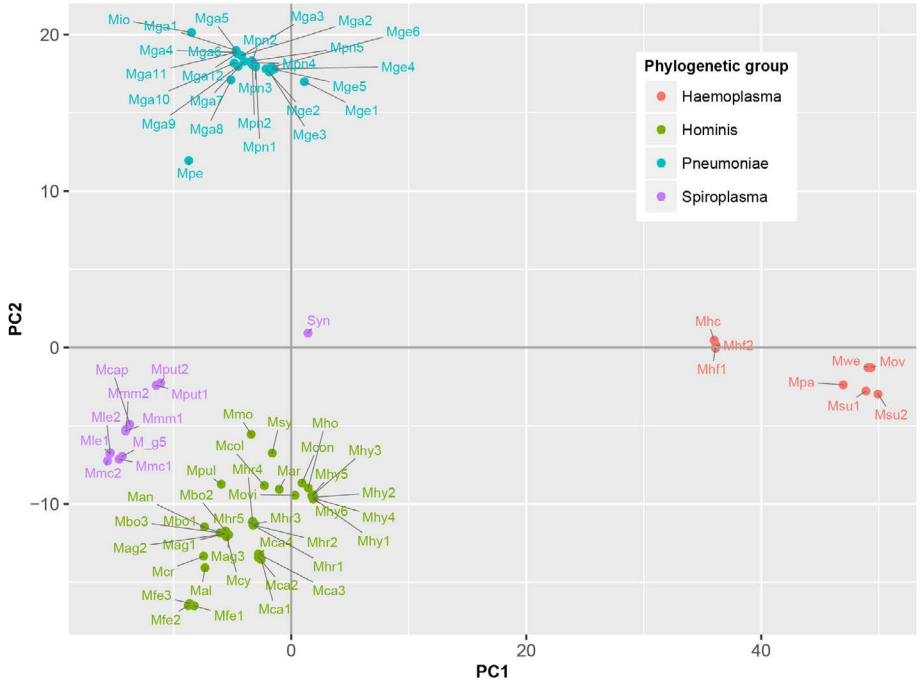


Fig. 4. Functional differentiation of mycoplasma species. Score plot is shown of principal component analysis done on the presence/absence matrix of the 80 mycoplasma strains and the synthetic bacterium JCVI-Syn3.0. Main phylogenetic groups are color coded. Note the separation of mycoplasma species infecting blood and tissue. Abbreviations used: Mag, *M. agalactiae*; Mal, *M. alligatoris*; Man, *M. anatis*; Mca, *M. canis*; Mcol, *M. columbinum*; M_g5, *M. g5847*, Mge, *M. genitalium*; Mio, *M. iowae*; Mmc, *M. mycoides capri*; Movi, *M. ovipneumoniae*; Mpn, *M. pneumoniae*; Mar, *M. arthritidis*; Mbo, *M. bovis*; Mca, *M. capricolum* subsp. *capricolum*; Mcon, *M. conjunctivae*; Mcr, *M. crocodyli*; Mcy, *M. cynos*; Mfe, *M. fermentans*; Mga, *M. gallisepticum*; Mhc, *M. haemocanis*; Mhf, *M. haemofelis*; Mho, *M. hominis*; Mhy, *M. hyopneumoniae*; Mhr, *M. hyorhinis*; Mle, *M. leachii*; Mmo, *M. mobile*; Mmm, *M. mycoides* subsp. *mycoides*; Mov, *M. ovis*; Mpa, *M. parvum*; Mpe, *M. penetrans*; Mpul, *M. pulmonis*; Mput, *M. putrefaciens*; Msu, *M. suis*; Msy, *M. synoviae*; Mwe, *M. wenyonii*; Syn, JCVI-Syn3.0. Numbers relate to strains (Table S4).

Functional differentiation between the hominis/spiroplasma and pneumoniae groups

Along the second principal component the hominis and spiroplasma clusters were separated from the pneumoniae cluster. We found 43 domains present in the hominis/spiroplasma clusters that mainly contributed to separation from the pneumoniae cluster versus 71 in the pneumoniae cluster that mainly contributed to separation from the hominis/spiroplasma

cluster (table 1 and table S5). In the hominis/spiroplasma cluster there was an increased presence of domains related to transport of magnesium and other divalent cations and also an increased capacity for chromate transport. Metals are important co-factors and increased chromate transport capability possibly results in increased chromate resistance as observed in *B. subtilis*[166]. Functionalities of other domains important to separate the hominis/spiroplasma cluster from the pneumoniae cluster were related to DNA/RNA modification, protein/peptide degradation and phosphopentomutase activity. The latter enzyme links nucleotide synthesis to the pentose phosphate pathway (PPP)[167] and provides mycoplasma with the option to produce nucleotides from the purine/pyrimidine bases or alternatively to degrade nucleotides via the PPP and glycolysis. In the set of domains that mainly contributed to separation of the pneumonia cluster from the hominis/spiroplasma cluster, a functional domain related to NAD kinase activity, needed for the production of NADP+, was found. Another domain was found linked to activity in the non-mevalonate pathway of isoprenoid synthesis: 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase. Activity of this pathway was shown for *M. penetrans* and *M. gallisepticum*[168] and might reduce the need to obtain isoprenoid precursors from the host. There was an increased presence of a domain related to thioredoxin-disulfide reductase activity which produces reduced thioredoxin needed for the production of deoxyribonucleotides and is important for protection against oxidative stress[169]. The separation of mycoplasma species based on protein domain composition provided a concise overview of the functional differences between mycoplasma species.

Table 1. Top 10 domains responsible for separation of mycoplasma functional clusters.

Enriched in haemoplasma[a]		Enriched in hominis/spiroplasma[b]		Enriched in pneumoniae[c]	
ID[d]	*InterPro description*[e]	*ID*[d]	*InterPro description*[e]	*ID*[d]	*InterPro description*[e]
IPR026023	Ribonucleotide reductase small subunit, prokaryotic	IPR029048	Heat shock protein 70kD, C-terminal domain	IPR002606	Riboflavin kinase, bacterial
IPR029022	ABC transporter, BtuC-like	IPR013826	DNA topoisomerase, type IA, central region, subdomain 3	IPR003526	2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase
IPR000522	ABC transporter, permease protein	IPR004398	RNA methyltransferase, RsmD	IPR006660	Arsenate reductase-like
IPR001674	GMP synthase, C-terminal	IPR003442	tRNA threonylcarbamoyl adenosine modification protein TsaE	IPR011631	Protein of unknown function DUF1600
IPR004837	Sodium/calcium exchanger membrane region	IPR006667	SLC41 divalent cation transporters, integral membrane domain	IPR023344	Uncharacterised domain MG237, C-terminal
IPR001670	Alcohol dehydrogenase, iron-type	IPR006668	Magnesium transporter, MgtE intracellular domain	IPR015271	Protein of unknown function DUF1951
IPR001093	IMP dehydrogenase/GMP reductase	IPR016947	Bacteriophage gamma, gammalsu0035	IPR013825	DNA topoisomerase, type IA, central region, subdomain 2
IPR019065	Restriction endonuclease, type II, NgoFVII	IPR000748	Pseudouridine synthase, RsuA/RluB/E/F	IPR012760	RNA polymerase sigma factor RpoD, C-terminal
IPR020471	Aldo/keto reductase subgroup	IPR001525	C-5 cytosine methyltransferase	IPR001844	Chaperonin Cpn60
IPR023210	NADP-dependent oxidoreductase domain	IPR003370	Chromate transporter	IPR002423	Chaperonin Cpn60/TCP-1

[a]Domains enriched in haemoplasma functional cluster

[b]Domains enriched in hominis/spiroplasma functional cluster

[c]Domains enriched in pneumoniae functional cluster

[d]InterPro Identifier

[e]Domain description obtained from InterProScan

Persistence of protein domains and of orthologous proteins

In order to compare the persistence of protein domains with the persistence of orthologous proteins, the complete set of orthologous proteins in the 80 mycoplasma genomes was determined using a standard bidirectional best hit approach[154] followed by orthology assessment with orthAgogue[155] and MCL clustering[156]. We found >5000 clusters of orthologous proteins and examined in how many genomes these orthologous proteins are present (figure 5A). Only 135 orthologous proteins are conserved amongst all mycoplasma species and we find an average persistence of orthologous proteins of 12.6%. The persistence of protein domains in the pan-domainome was much higher (average of 48.4%, figure 5A).

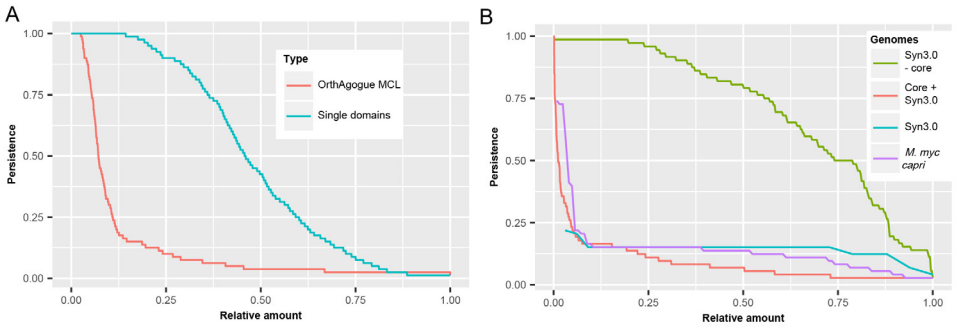


Fig. 5. Persistence of (essential) domains and (essential) hypothetical proteins. A: Persistence of orthologous proteins based on sequence similarity (red) and persistence of single domains (blue). B: Persistence of domains present in JCVI but absent from the core of the mycoplasma species infecting tissue (green). Persistence of hypothetical proteins for 72 mycoplasma species including JCVI-Syn3.0 (red), *M. mycoides capri* LC (purple) and JCVI-Syn3.0 (blue).

Mycoplasma pan- and core-domainome analysis in relation to JCVI-Syn3.0

Clustering of mycoplasma species based on the pan-domainome did not show a correlation with their specific host. To further classify protein domains, we compared the pan- and core-domainome of the mycoplasma genus with the domainome of the minimal synthetic organism JCVI-Syn3.0[86] which consisted of 869 domains (table S7). For the synthetic organism we assumed that all protein domains in this organism were essential. The core domainome consisted of 335 domains, a relatively small amount. This can be explained because we took into account the haemoplasma species that grow in blood and cannot be cultured *ex vivo*. When the core-domainome was calculated for mycoplasma species infecting tissue, a larger size core was obtained of 479 protein domains. From this core 26 domains were not present in JCVI-Syn3.0 (table S8). Apparently these domains are not essential for growth in a laboratory environment. The remaining 453 core domains show overlap with JCVI-Syn3.0 (table S8), indicating that these persistent domains are essential for axenic growth in (complex) growth media. Interestingly, the remaining 416 domains in

JCVI-Syn3.0 essential for minimal life are not persistent (figure 5B and figure S3) suggesting that within the mycoplasma domain landscape many alternative configurations exist that bypass their essentiality.

Metabolic capability in relation to host specificity

To assess if domains with a non-essential metabolic function determine host specificity, we obtained all domains with a metabolic function not present in the synthetic minimal organism. Domains with a metabolic function were derived based on the genome-scale metabolic model of *M. pneumoniae*[67] supplemented with InterPro annotations. This model contains 145 genes, coding for 145 proteins, and from this set of proteins 359 unique protein domains were obtained. Almost all proteins with a metabolic function were covered with domains (97%). Overall we found 162 domains (33.8% of the total core) with a metabolic function to be present in the core of the tissue infecting mycoplasma species and 197 accessory domains with a metabolic function present in the accessory domainome (pan minus core). In JCVI-Syn3 156 domains from the metabolic core domainome and 140 accessory domains with a metabolic function were present. Thus 63 domains with a metabolic function were absent in JCVI-Syn3.0 and to assess whether these domains could be involved in host specificity we clustered mycoplasma species infecting tissue based on the presence/absence of these domains but we could not establish a correlation with host specificity (figure S4).

Role of hypothetical proteins in host adaptation

Clustering based on the pan-domainome composition or on the metabolic domain complement absent in JCVI-Syn3.0 failed to show a direct link between specific domains and host specificity. We further analyzed if presence or absence of hypothetical proteins could explain host specificity. In our dataset, a protein was annotated as hypothetical when a protein did not contain a protein domain or when a protein contained a domain of unknown function (DUF). In total 58 DUFs were found in the mycoplasma genus, from which only 8 DUFs were present in JCVI-Syn3 (table S9). There were no DUFs in the core domainome of the complete genus and only 2 DUFs in the core domainome of the tissue infecting mycoplasma species (DUF161 and DUF933). DUF161 is part of a membrane protein with unknown function; DUF933 is suggested to be part of a nucleoprotein complex and could function as a GTP-dependent translation factor. The total amount of DUFs found was too low to analyze a relation with the host and for further classification of hypothetical proteins we compared the complete set of hypothetical proteins in JCVI-Syn3.0 to the complete set of hypothetical proteins in the pan-genome of the mycoplasma species infecting tissue. In total 11,598 hypothetical proteins were found in the tissue infecting mycoplasma species which based on sequence similarity, could be clustered into 1,766 orthologous protein clusters. The relative persistence of the hypothetical protein clusters showed a sharp decline with an average persistence of approximately 9% (figure 5B). The

total amount of genes with completely unknown functions in the genome of JCVI-Syn3.0 was only 65 [86] and we identified just 40 proteins to which no functional domains could be assigned. The persistence of orthologous protein clusters containing these hypothetical proteins was 14% (figure 5B) which was higher than average. There was, however, conservation of clusters with hypothetical proteins from the spiroplasma phylogenetic group. In line with the finding that not all essential JCVI-Syn3.0 protein domains were persistent, essential hypothetical proteins were also not persistent suggesting that within the mycoplasma genus alternative solutions exist substituting these essential but currently unknown functions. We did not observe a relation with the host on the basis of the clustering of orthologous hypothetical proteins not present in JCVI-Syn3.0 (Fig. S5).

Protein domain composition in relation to host or niche

Clustering based on the complete pan-domainome of mycoplasma, the metabolic domains outside JCVI-syn3.0 as well as the hypothetical orthologous proteins outside JCVI-Syn3.0 did not show a relation with a mycoplasma species specific host. As a final effort, we applied two machine learning approaches: k-nearest neighbor (k-nn) and Random Forest[170], to classify a mycoplasma species niche or host based on the pan-domainome composition. Both methods could predict with high accuracy whether the niche of a mycoplasma species is blood or tissue confirming the results already found using PCA (supplementary materials and table S5). When the niche was specified in more detail (table S6, Niche), the prediction accuracy decreased and species with a unique niche (e.g. *M. mobile* and *M. conjunctivae*) could not be assigned. Classification of mycoplasma growing in blood, strictly in the respiratory tract and in multiple tissue types including lung (table S6, Niche 2) was possible using Random Forest with 95% prediction accuracy (5% out-of-bag error rate). The domain most important for classification was cell division protein *FtsZ* (IPR000158). This domain was present in many mycoplasma species but absent from *M. canis*, *M. gallisepticum* and *M. hyopneumoniae*, which formed for a large part the species infecting the respiratory tract in our dataset. Absence of this specific domain does not mean that a species has no functional *FtsZ*, since there are alternative domain configurations possible (containing e.g. domain IPR003008 and IPR020805). To prevent prediction bias due to differences in the number of genomes available of a certain species, we decided to focus on the mycoplasma species infecting tissue for which we had at least two genomes and limited our search to two genomes per species. Using this smaller selection of genomes, prediction accuracy was higher (96% using the random forest classifier and 71% using k-nn classification) and we again identified the specific *FtsZ* domain (table 2 and table S10) as an important domain for niche classification. We also identified a putative DNA-binding domain (IPR009061), present in phenylalanine-tRNA synthetases. In our database this domain was not present in the selected strains of *M. canis*, *M. hyopneumoniae* and *M. pneumoniae* which are all present strictly in the respiratory tract. The domain was, however, present in other mycoplasma species identified as strictly present in the respiratory tract: *M. cynos*, *M.*

gallisepticum and *M. mycoides* subsp. *mycoides* SC. Also important for classification was restriction endonuclease, type I domain IPR000055, which was not present in *M. gallisepticum* strains used in our selection and was also absent from the *M. mycoides* subsp. *mycoides* SC strains used in our comparison. There was not a single domain uniquely present in all mycoplasma infecting the respiratory tract and absent from the mycoplasma infecting multiple tissue types.

Table 2. Top 10 domains relevant for niche classification: strictly respiratory or multiple tissue types.

Domain information		Abundance (%)[a]	
ID[d]	*InterPro description*[e]	*Respiratory system*[b]	*Multiple*[c]
IPR009061	DNA binding domain, putative	40	100
IPR000055	Restriction endonuclease, type I, HsdS	50	94
IPR022749	N6 adenine-specific DNA methyltransferase, N12 class, N-terminal	20	81
IPR000158	Cell division protein FtsZ	40	100
IPR011701	Major facilitator superfamily	50	88
IPR003798	DNA recombination RmuC	20	75
IPR008280	Tubulin/FtsZ, C-terminal	40	88
IPR002198	Short-chain dehydrogenase/reductase SDR	20	75
IPR011089	Domain of unknown function DUF1524	0	50
IPR005864	ATPase, F0 complex, subunit B, bacterial	60	100

[a]Abundance of a protein domain in the specific niche
[b]Abundance in mycoplasma species with a strictly respiratory niche
[c]Abundance in mycoplasma species with multiple niches including respiratory
[d]InterPro Identifier
[e]Domain description obtained from InterProScan

For identification of domains important to classify mycoplasma hosts, we first used the complete diversity in hosts mentioned in table S6 and obtained a prediction accuracy of <80% using random forest. We decided to use a more focused approach and selected only mycoplasma species growing in tissue for which we had two species per host and two genomes per species. Genomes for cows and goats were pooled into a ruminants group. With this grouping, we could accurately predict (83% accuracy with k-nn classification and 100% with random forest) if a mycoplasma species from the selected genomes infects a pig, ruminant or human. The most discriminatory domains identified from the random forest analysis (table 3 and table S11) were related to peptidase functions (IPR000668, IPR005151 and IPR029045). A phosphodiesterase domain (IPR024654 and related family IPR000979) was found to be important for host differentiation, this domain only occurs in

the human pathogens taken into account. A *RmlC*-like jelly roll fold domain (IPR014710), which is related to mannose/myo-inositol metabolism, was identified in the pig and ruminant species but was absent from species that infect humans. Two domains of unknown function were found: DUF2714 and DUF285 (IPR021222 and IPR005046). The DUF285 domain has probably been exchanged between ruminant species via horizontal gene transfer[171]. Several domains related to proteins expressed at the bacterial surface were found (IPR011889 and IPR027593). A glycine cleavage domain was found (IPR002930) which was absent from the selected mycoplasma species infecting humans. Using the Random Forest prediction, on specific species groups, we have identified a number of protein domains which could relate to host specificity.

Table 3. Top 10 domains relevant for host classification: ruminants, pigs or humans.

Domain information		Abundance (%)[a]		
ID[e]	*InterPro description*[f]	*Ruminants*[b]	*Pigs*[c]	*Humans*[d]
IPR000668	Peptidase C1A, papain C-terminal	100	0	0
IPR005151	Tail specific protease	100	0	0
IPR000979	Phosphodiesterase MJ0936/Vps29	0	0	100
IPR014710	RmlC-like jelly roll fold	100	100	0
IPR021222	Protein of unknown function DUF2714	100	100	0
IPR005046	Protein of unknown function DUF285	100	0	0
IPR002931	Transglutaminase-like	100	0	0
IPR002930	Glycine cleavage H-protein	100	100	0
IPR011889	Bacterial surface protein 26-residue repeat	92	0	0
IPR029045	ClpP/crotonase-like domain	100	0	0

[a]Abundance of a protein domain in the specific host
[b]Abundance in ruminant species
[c]Abundance in pig species
[d]Abundance in humans
[e]InterPro Identifier
[f]Domain description obtained from InterProScan

Discussion

All mycoplasma species have reduced genomes and could be considered minimal organisms. Arguably, the most studied minimal organism to date is *Mycoplasma pneumoniae*, a human pathogen causing inflammation of lung tissue in humans. From this bacterium we have knowledge of the genome[172], transcriptome[55], proteome[74], metabolome[75,173] and several regulatory mechanisms including the role of non-coding RNA's (chapter 6). Interactions with the host have also been extensively studied[48] but

despite of the wealth of information on this minimal organism we still cannot explain why there is a preference for colonization of human lung tissue. Knowing that it is not a simple case of adhesion properties[174], we hypothesized that there is a complex combination of functions that determines a bacterial host or niche. To find these functions, we clustered species based on domain presence to find direct leads and ultimately used a random forest classification algorithm on the complete mycoplasma pan-domainome to find sets of domains that predict the specific host or niche of a mycoplasma species. By considering presence or absence of proteins domains we deviate from the classical approach in which bacterial genomes are compared based on orthologous proteins. We found that the persistence of single domains is higher which indicates that conservation of the structural information in the protein domains is more important than maintaining the gene sequences in which the domains are present. A similar result was recently found in a comparative genomics study of 432 *Pseudomonas* species[59], indicating that this could be a trend amongst bacterial species. By using Random Forest classification, we could predict with high accuracy whether a mycoplasma species infects tissue or blood and found metabolic properties in the haemoplasma cluster that could explain why this organism successfully infects blood. Zooming into functional species clusters, the prediction accuracy decreases and it is not possible to predict a host or niche of closely related species such as *M. haemofelis* and *M. haemocanis,* within the haemoplasma group, or *M. agalactiae* and *M. bovis*, within the hominis group. Despite the lower prediction accuracy we were still able to identify differences between mycoplasma species in relation with its specific host or niche if we used larger clusters as was shown for the differentiation of mycoplasma colonizing ruminants, pigs or humans. To determine the specific role of a signifying protein function (e.g. one of the peptidase functions) in host-pathogen interaction would require additional laboratory studies.

To understand in greater detail which factors determine host or niche specificity, more mycoplasma genomes of species of specific interest could be sequenced. This will provide more detailed information on the variation in the domain composition of this species, increasing the accuracy of host prediction. Further information needed to understand host or niche specificity could also follow from functional annotation of proteins without a protein domain, which make up approximately 20% of the total proteome of a mycoplasma species. The machine learning approaches applied did not take domain abundance into account as we used the binary domain matrix as input to avoid overfitting. (Dual-)Trancriptomics studies might provide the additional insight needed to explain the interplay between host and pathogen. For example, a recent study on the chicken pathogen *M. gallisepticum*[175] showed temporal phase variation in the expression of *vlhA* genes during infection. Finally, the strict host specificity for mycoplasma species can be challenged since several mycoplasma species infect a broad range of hosts (e.g. *M. bovis* and *M. mycoides* subsp. *mycoides*) and mycoplasmas normally isolated from animals are sometimes found in humans and vice versa[176,177]. The assumption of strict host specificity

for mycoplasma species could be incorrect and mycoplasma may be able to infect a wider range of hosts and ecosystems than previously anticipated[72].

Our finding that the pan-domainome of the mycoplasma genus is closed supports the general expectation that species with an allopatric lifestyle have a lower chance of gaining genes by horizontal gene transfer (HGT). This finding, however, seems to contradict the recent comparative genomics reports on an open pan-genome for mycoplasma species[143,144]. Possible mechanisms that could contribute to the increase of the pan-genome have been described to be: 1) variation in expression and structure of surface antigens, 2) horizontal gene transfer (HGT), 3) genetic drift and 4) phage attack[72]. HGT events between species outside the mycoplasma genus are rare[70] and phage attacks are not common in mycoplasma species[178]. Thus, we expect that genetic drift and sequence variations in the regions coding for variable surface proteins contribute to an increase in the pan-genome size but that this increase is mainly related to genes encoding proteins without characterized domains.

Because the pan-domainome of mycoplasma species is closed, sequencing additional strains will not add to the overall systems level understanding of mycoplasma physiology and focus should be on further understanding of the mycoplasma strains for which the genome sequence is known. In this study we incorporated the minimal JCVI-Syn3.0, which is based on a *M. mycoides* template. We considered a protein domain essential when it was present in the minimal synthetic bacterium meaning that the protein domain is needed for growth in a complex cultivation medium under laboratory conditions. We also consider it likely that none of the domains in the minimal synthetic bacterium are needed to maintain growth in the specific host since the genome has been minimized for growth outside the host. By comparing the core domainome of the mycoplasma genus with JCVI-Syn3.0 we found that almost all domains present in the mycoplasma core are also present in the minimal synthetic organism and are likely needed to support growth in axenic media under laboratory conditions. The synthetic bacterial genome still contains 17% of essential protein coding genes with an unknown function. We found that conserved hypothetical proteins in the spiroplasma functional group are conserved in JCVI-Syn3.0. This finding is in line with the general notion that conserved hypothetical proteins are more likely to be essential[179] but in the case of mycoplasma this conservation is limited to mainly the functional cluster, and not to the complete genus. Both findings can provide a guideline for the design of minimal bacterial synthetic genomes. We expect that when mycoplasma species from other functional groups are taken as a template, alternative configurations will emerge showing flexibility in the composition of the pan-domainome of minimal synthetic bacteria designed from mycoplasma ancestors.

Acknowledgments

We thank Pascal Sirand-Pugnet for critically reviewing the manuscript.

Chapter 5

Transcriptome sequencing shows up-regulation of F_1-like ATPase and down-regulation of the P102 cilium adhesin in *Mycoplasma hyopneumoniae* during infection

Tjerko Kamminga, Vitor Martins dos Santos, Jetta J.E. Bijlsma and Peter J. Schaap

Manuscript prepared for submission

Abstract

Mycoplasma hyopneumoniae causes enzootic pneumonia in pigs. The exact mechanisms used by *M. hyopneumoniae* to colonize and survive in the pig lung are not completely understood. Insight into the infection process can be obtained by deep-sequencing of the bacterial transcriptome during infection but isolation of sufficient bacterial mRNA from infected tissue is challenging due to the high concentration of host RNA. In the current study we optimized a sampling protocol to obtain sufficient bacterial mRNA from infected lung tissue for RNA sequencing. We compared the transcriptional landscapes of *M. hyopneumoniae* in four biological replicates to duplicate *in vitro* cultures and identified 22 up-regulated and 30 down-regulated genes (FDR<0.01 and fold change >2LOG2). Six out of seven genes in the operon encoding the mycoplasma specific F_1-like ATPase (MHP_RS02445-MHP_RS02475) were up-regulated *in vivo* and all genes in the operon MHP_RS01965-MHP_RS01990 with functions related to nucleotide metabolism, spermidine transport and glycerol-3-phoshate transport were up-regulated *in vivo*. Down-regulated *in vivo* were genes related to glycerol uptake, cilium adhesion (P102), cell division and myo-inositol metabolism. Besides providing a novel method to isolate bacterial mRNA from infected lung, this study provided insight into gene expression during infection which could help development of novel treatment strategies against *M. hyopneumoniae* infection in pigs.

Introduction

Mycoplasma hyopneumoniae causes enzootic pneumonia in pigs[180], a mild, chronic pneumonia characterized by a non-productive, dry cough. Infected pigs often develop secondary infections which makes *M. hyopneumoniae* an important contributor to the development of respiratory disease complex in pigs[2] and a major threat to the worldwide pig industry. The pathogen easily spreads within and between herd populations via nose-to-nose contact and aerosols. Treatment with antibiotics results in a decrease in symptoms but not in eradication of the disease. Vaccination with adjuvanted inactivated bacterial vaccines is effective to control disease symptoms but does not prevent colonization of the lung. Improvement of housing conditions and herd management practices can also decrease disease prevalence[5]. There is a need for improved treatment or prevention methods but development of these methods is hampered because the exact mechanisms used to colonize and survive in the pig lung are not completely known[17,181].

Investigations into *M. hyopneumoniae*-host interactions using histological techniques have shown that this pathogen specifically colonizes the lower respiratory tract by binding to cilia on the epithelium of trachea, bronchi and bronchioles. The entire length of the cilia can be bound by bacteria but interaction with the epithelial cell body is rare[21,35,40]. *M. hyopneumoniae* is not known to be motile and does not form motility structures, such as those found in *M. mobile* and *M. pneumoniae*, yet the bacterium successfully evades the protective action of the mucociliary escalator. *M. hyopneumoniae* expresses a large diversity of proteolytically-cleaved multifunctional cilium adhesion proteins on the bacterial cell surface, which bind glycosaminoglycans (e.g. heparin), fibronectin and plasminogen[27,30,182,183]. Most abundantly present on the cell surface are fragments from P97 and P102 proteins for which genes are present in paralogous gene families with six copies per gene in the *M. hyopneumoniae* strain 232 genome[15]. Many of these paralogs are present in two gene transcriptional units with one copy per gene and were found to be expressed *in vivo*[184]. Besides the P97/P102 paralogous families, two other *M. hyopneumoniae* genes function as cilium adhesins: P159 (MHP_RS02535) which is proteolytically cleaved into three fragments that bind heparin[33] and M42 glutamyl aminopeptidase (MHP252) which binds heparin and plasminogen[34]. The repertoire of proteins used for adhesion is probably even more extensive since many intracellular proteins were also found to be present on the cell surface suggesting a possible role in adhesion[185].

When attached to the ciliated epithelium, multiple bacterial lipases, proteases and nucleases could release nutrients for growth but specific virulence factors have not been described. Multiple lipoproteins (P65, P50, P44 and P70) are expressed at the bacterial cell surface and were found to be highly immunogenic. P65 (MHP_RS03425) was found to be a lipolytic enzyme with a preference for short-chain fatty acids[47]. Upstream of the P65 gene lies a region with tandem repeats which is expected to cause slippage of DNA polymerase

which could cause variation in the expression of the P65 protein[16,186]. P65 has two paralogs in the *M. hyopneumoniae* strain 232 genome (MHP_RS02755 and MHP_RS00345)[15]. Functions for the other lipoproteins are unknown. The recruitment of plasminogen and activation to plasmin, facilitated by the M42 glutamyl aminopeptidase, is a potential mechanism that could cause tissue damage[34,38]. Further damage to host tissue could be caused by the production of oxidizing compounds, such as hydrogen peroxide or hydrogen sulfide, as was reported for other mycoplasma species[36,140,187,188] but whether this mechanism plays a role in *M. hyopneumoniae* infections remains to be established. Finally, differences in the metabolic capabilities between commensal and pathogenic swine mycoplasma species have been elucidated using genome-scale metabolic models[116]. In this study the glycerol pathway, related to hydrogen peroxide production and the myo-inositol pathway, uniquely present in *M. hyopneumoniae* when compared to other mycoplasma species, were reported as possible pathways related to virulence. To further understand the role of the adhesive proteins, possible virulence factors and metabolic pathways during infection, the bacterial *in vivo* transcriptome needs to be determined.

Bacterial gene expression during infection has been studied *in vivo* using microarrays by Madsen *et al.*[189]. In this study, glycerol metabolism was found to be up-regulated *in vivo* supporting the possible role of glycerol oxidase during infection. Determination of the transcriptome of pathogens with microarrays has several disadvantages, such as: probe sequence bias, non-specific hybridization and the dynamic range in expression levels that can be measured is low since the detection method has a saturation point[190]. Because of these disadvantages, microarrays have been widely replaced by RNA-sequencing. Studying bacterial gene expression during infection with RNA sequencing is challenging because the ratio of host RNA to bacterial RNA is very high in infected tissue which makes it hard to obtain sufficient bacterial read numbers to assign expression levels for genes. So far, only one study analyzed gene expression in a mycoplasma species using RNA sequencing but this study did not give a global insight in the transcriptional landscape of the bacterium inside the host. Therefore, we developed a sampling method to isolate bacterial RNA from *M. hyopneumoniae* infected lung tissue and we applied RNA sequencing to analyze the transcriptome of *M. hyopneumoniae* during *in vivo* growth and compared it to *in vitro* gene expression. This novel insight into differential gene expression during infection could provide important leads for development of novel treatment strategies.

Materials and methods

Bacterial cultivation, DNA isolation, RNA isolation and genome sequencing

Mycoplasma hyopneumoniae strain 98, a Danish field isolate strain (provided by Dr N. Friis, National Veterinary Laboratory, Copenhagen) was used. The strain was grown in 100 ml FRIIS medium[117] in 250 ml closed glass bottles at 37°C with agitation (100 RPM).

Cultures were sampled during exponential growth indicated by an increase in titer measured using flow cytometry (FACSMicroCount, BD). Mycoplasma cells were pelleted using centrifugation (3 minutes at 9000 g's) and the cell pellet was directly frozen (<-15°C) for DNA extraction or submerged in RNA later (Ambion) and stored at 2-8°C for RNA extraction. DNA extraction for genome sequencing from the pelleted frozen cells was done using the Gentra Puregene bacterial kit (Qiagen). Genome sequencing was done using Illumina HiSeq2500 (paired-end, 100 cycles, 500 mb, 150 bp read length) as previously described in chapter 3. Genome assembly was done using the Ray algorithm[191]. Annotation of the genome was done using SAPP[59] in which Prodigal v.2.6.2[120] was used for gene calling and InterProScan v.5.17-56.0[58] for protein domain annotation. A reciprocal best-blast hit analysis[154] was done using BLAST+[192] to find orthologous proteins in *M. hyopneumoniae* strain 232[15] filtering for matches with a minimum of 70% identity and 50% query coverage.

Animal studies

Animal studies were performed after approval by an ethical commission and according to national regulations in The Netherlands. Six healthy pigs from a *M. hyopneumoniae* free herd were challenged intra-tracheally at seven weeks of age on two consecutive days with 10 ml of *M. hyopneumoniae* strain 98 culture containing $\pm 10^7$ CCU/ml. Pigs were anesthetized, euthanized and exsanguinated three weeks after challenge. Within a short time-period (<5 minutes) after the necropsy the lungs were removed and tissue samples were obtained and directly submerged in RNA later (Ambion) or infected lobes were flushed with RNA later using a method previously described[193] but adapted for this specific purpose. Briefly, an infected lobe was selected for sampling based on the presence of lesions in the lobe. A plastic pipette (7 ml total volume) was used to dispense 5 ml of RNA later (Ambion) into the bronchus towards the selected lobe. Directly after dispensing, the fluid was retrieved from the lobe and stored at 2-8°C for further processing.

RNA extraction

RNA extraction from tissue samples was done using the RNeasy mini kit (Qiagen) following the manufacturer's protocol. RNA extraction from flush samples (0.75 ml) was done with 7.5 ml of Trizol LS reagent (Ambion) following the manufacturer's protocol. For bacterial pellets, RNA later was removed and 3.75 ml Trizol LS reagent was added. RNA was precipitated using glycogen as co-precipitant as described in the Trizol extraction protocol. After RNA re-suspension, the quantity was determined using Nanodrop. RNA samples were treated with Turbo DNase (Ambion) to remove potential DNA contamination following the manufacturer's protocol. Enrichment of bacterial RNA in the flush samples and culture samples was done using the MICROBEnrich kit (Ambion) and the quality of all samples was determined using Experion RNA StdSense (Bio-Rad).

RNA sequencing and read mapping

rRNA was removed using the Ribo-Zero Gold rRNA Removal Kit (Epidemiology). rRNA-depleted RNA was fragmented to an average length of 100 to 200 base pairs and converted to double-stranded complementary DNA (cDNA). Strand specific library preparation was done using a protocol based on the "dUTP (deoxyuridine triphosphate) method". The Illumina stranded TruSeq RNA-seq library preparation kit was used. Sequencing of the library was done using the Illumina HiSeq 2500: single-end reads, 50 cycles. Quality assessment of reads was done as previously described in chapter 3. Reads were aligned to the genome using TopHat v2.1.0[194] with the standard settings for strand-specific alignment enabled (fr-firststrand). Accepted hits found by TopHat were sorted using samtools v1.3.1[195], from the sorted hits a BED-file was created using bamToBed and coverage per gene was calculated using coverageBed, both programs from the bedtools suite v2.17.0[196]. Reads mapping to non-coding RNAs, were determined using the same methods after creation of a ncRNA annotation file for the strain 98 genome based on the best blast hit[192] with the annotated ncRNAs in strain 232 (chapter 6).

Analysis of differentially expressed genes and differentially expressed ncRNAs

The R package edgeR[197,198] was used for analysis of differentially expressed genes and differentially expressed ncRNAs (for simplicity we refer to both as genes) between the *in vitro* and *in vivo* condition. Only genes with a read count of >100 counts per million (CPM) in two or more datasets were kept as described by Rienksma *et al.*[199]. Data was normalized for RNA composition by calculating scale factors (based on the trimmed mean of M-values[200]) which were used to correct the library size. Common dispersion was calculated based on conditional maximum likelihood, correcting for library size differences based on pseudocounts estimated per quantile, first by assuming a poisson distribution to calculate the common estimated dispersion. This initially calculated common dispersion was used to get a final estimate for the pseudocounts and a final estimate of the common dispersion[201]. Two types of variation contribute to the total variance in the probability distribution: technical and biological variation. Biological variation (BCV) represents the variation between biological replicates that would remain if sequencing depth is infinite. Dispersion per gene (biological variation) was calculated using an empirical Bayes estimation with weighted likelihood[202]. Differential expression was calculated for each gene using a pairwise exact testing method taking into account the tag-wise dispersion estimates. False discovery rates were controlled using the algorithm by Benjamini and Hochberg [203]. Genes with a FDR<0.01 and a LOG2 fold change larger than 2 were considered to be differentially expressed.

Results and discussion

Properties of the *Mycoplasma hyopneumoniae* strain 98 genome

Genome assembly using the filtered Illumina FastQ sequence reads (Phred quality score 36.17) resulted in a draft genome sequence of 19 scaffolds. The total genome size was 880.620 bp and the GC% was 28.5%. We compared the protein domain content of strain 98 to the domain content of all other sequenced *Mycoplasma hyopneumoniae* strains (232, J, 11, 7448, 7422, 168 and 168L, table S1) obtained after genome annotation using SAPP[59] as described in chapter 4. We found a total core domainome size of 843 unique protein domains and an accessory domainome of 23 protein domains (table S2). We found no protein domains in strain 98 that were unique for this strain. Two protein domains were absent from strain 98 although they were present in all other *M. hyopneumoniae* strains: IPR004372, "Acetate/propionate kinase", and IPR004541, "Translation elongation factor EFTu/EF1A, bacterial/organelle", respectively found in acetate kinase (MHP_RS02595) and elongation factor Tu (MHP_RS02760). Homologous genes were identified in the strain 98 genome by reciprocal best blast hit analysis and the respective proteins contained protein domains with functional annotations matching the functions in the other *M. hyopneumoniae* strain, showing that the protein functionality is equivalent but the domain configuration is slightly different. IPR000209, "Peptidase S8/S53 domain", was also uniquely absent from strain 98 and a reciprocal best blast analysis with MHP_RS01485, the gene in strain 232 that contains this domain, did not result in a hit indicating that this gene is absent from strain 98. In total, with a reciprocal-best blast hit analysis 601 orthologous proteins were identified in strain 98 when compared to strain 232. Based on the composition of the accessory domainome, strain 98 is most related to *M. hyopneumoniae* strain 7448 (figure S1). Overall, the differences between *M. hyopneumoniae* strains were small and no novel protein domains were identified in the strain 98 genome. The genome sequence of strain 98 was used for mapping of reads obtained from RNA sequencing, for interpretation of the results we mention the locus tags of the homologous proteins in strain 232.

Optimized method to obtain sufficient bacterial mRNA from infected lungs

In this study we tested various methods to isolate bacterial RNA from lungs infected with *M. hyopneumoniae* (figure 1a). Lesions in infected lungs were identified by discoloration of tissue and from the affected tissue we extracted total RNA for sequencing. However, we expected a high concentration of pig RNA in these extracts and investigated methods to more specifically isolate bacteria from the infected lung. Application of differential lysis techniques to isolate bacterial RNA[199,204], was not an option for *M. hyopneumoniae* since this bacterium does not contain a cell wall. As previously reported, we identified with immunohistochemistry that *M. hyopneumoniae* is found specifically on the surface of the respiratory epithelium (figure 1a and chapter 1). To isolate a maximum amount of bacteria,

we decided to flush an infected lung lobe with RNA later and isolate total RNA from the lung flush. However, since we still expected host cells to be present in the lung flush, with a high concentration of RNA compared to bacteria, we also applied an enrichment step for bacterial RNA in which pig mRNA is removed based on the presence of a poly-A tail (figure 1a). Quality of all purified RNA samples was analyzed on RNA gels which showed that the average total amount of RNA isolated with our flush protocol was low 4.2 µg (range: 0.17-18.51 µg). Due to the low quantity of RNA isolated from the flush samples it was not possible to assess the RNA quality based on the ratio between 23S and 16S rRNA (figures S2-S6). Therefore, we analyzed the quality of the *in vivo* data with multivariate data analysis. As a rule of thumb, we assumed that sufficient coverage of a bacterial genome for analysis of gene expression levels is obtained with ≥1 million non-rRNA bacterial reads[205]. Sequencing of mRNA isolated from infected tissue resulted in only 0.1% read mapping (figure 1b, table 1) to the bacterial genome and the sequencing capacity needed to reach sufficient reads using this sample type was considered not economically feasible. A ten times higher read mapping percentage was achieved by sequencing RNA from a flush sample and even higher read mapping percentages were reached by applying the enrichment step (figure 1b, table 1).

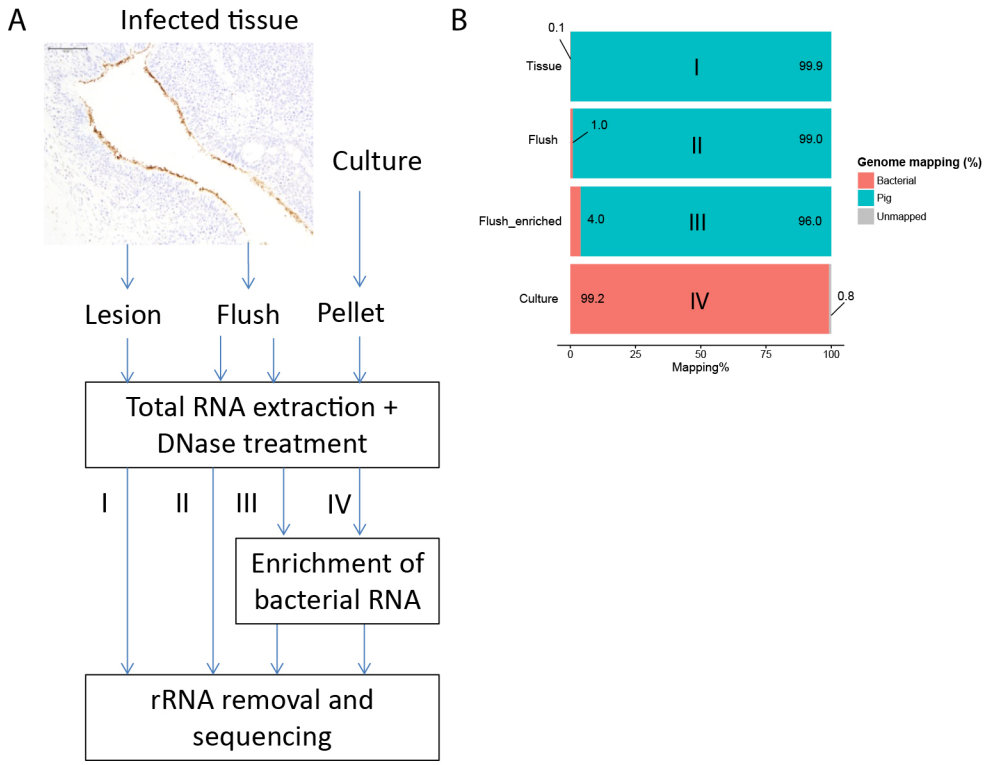


Figure 1. Development of an enrichment strategy to obtain bacterial RNA reads from infected tissue. Panel A: Immunohistochemistry showed that *M. hyopneumoniae* (stained brown) is bound to the respiratory epithelium. Three alternative strategies were tested; I: total RNA extraction from infected tissue; II: RNA extraction from lung flush; III: RNA extraction from lung flush with bacterial enrichment step, IV: *in vitro* control. B: Percentage of reads mapped to the *M. hyopneumoniae* genome.

To assess if the enrichment step influenced the expression levels for bacterial genes, we sequenced a flush sample from the same pig with and without applying the enrichment step, all other treatment steps were done in parallel. There was a strong positive correlation between the expression levels in the non-enriched sample compared to the enriched sample (Pearson coefficient = 0.97, figure S7) which showed that the enrichment step and also the repeat of the other purification steps did not influence the bacterial read distribution. We further analyzed the variation in the read distribution per bacterial gene in the various sample types with principal component analysis (PCA) based on the read counts per gene. This analysis showed a separation of *in vitro* samples from *in vivo* samples and again showed that enrichment for bacterial RNA did not influence the general expression landscape in biological replicates (figure S8). This analysis also showed that 69% of the variation in the bacterial read distribution was a result of the environment from which the samples were obtained and only 19% was caused by variation between *in vivo* samples,

showing high consistency for both sample types. The method we developed enabled sequencing of bacterial mRNA from infected lung tissue with sufficient read numbers.

Table 1. Read distribution for study samples. Quality filtered reads were mapped to the *Mycoplasma hyopneumoniae* strain 98 genome.

Origin	Bacterial culture		*M. hyo* infected lung tissue[a]		Flush *M. hyo* infected lung tissue[b]				
Sample name	CUL1	CUL2	T1	T2	F1-NE	F1-E	F2-E	F3-E	F4-E
Enrichment bacterial RNA[c]	Yes	Yes	No	No	No	Yes	Yes	Yes	Yes
Total number of reads (x million)[d]	24.3	31.9	75.3	68.3	66.6	37.3	94.5	76.9	38.5
Nr. reads mapping strain 98 genome (x million)[e]	24.1	31.6	0.1	0.1	0.5	1.4	2.9	1.9	2.6
% reads mapping strain 98 genome[f]	99.2	99.1	0.1	0.1	0.8	3.7	3.0	2.4	6.8

[a]RNA from infected tissue was sequenced, selected based on visual appearance (tissue discoloration)

[b]Infected lobes in the lung were flushed with RNA later

[c]Specification whether RNA enrichment was done using the MicrobeEnrich kit

[d]Total number of reads obtained after quality filtering

[e]Total amount of reads mapping to the strain 98 genome

[f]Percentage of reads mapping the strain 98 genome (mapping reads/total reads x 100%)

Highly expressed genes during infection were similar to highly expressed genes in culture

Genes needed for cilium adhesion were highly expressed during infection (table 2). This was expected because the bacterium specifically binds the ciliated epithelium (figure 1a). We found the same genes highly expressed *in vitro* which showed that there is no on/off switching of expression of these genes depending on the environment. Genes related to cell division and transport, specifically a subunit of the ascorbate transporter and MFS transporter, were also found highly expressed in both conditions. Key metabolic genes found highly expressed under both conditions were related to the final stages of glycolysis (pyruvate dehydrogenase and lactate dehydrogenase) and myo-inositol metabolism. This indicates an important role for these metabolic pathways under both growth conditions. Our results agreed well with data reported by Siqueira *et al*[54] of highly expressed genes in *M. hyopneumoniae* strain 7448 determined with RNA sequencing.

Table 2. Highly expressed genes in *M. hyopneumoniae in vivo* compared to highly expressed genes *in vitro*.

#[a]	Gene[b]	Strain 232 locus tag[c]	Average expression flush (RPKM)[d]	Average expression culture (RPKM)[e]	# cult.[f]	Strain 232 functional annotation[g]
1	3_47	02535	493776.2	299689.9	4	P97 paralog (P216)
2	2_38	02100	247739.6	734005.1	1	ribosomal RNA small subunit methyltransferase H
3	2_39	02105	225848.4	592086.0	2	cell division protein FtsZ
4	2_37	02095	114276.6	331599.7	3	transcriptional regulator MraZ
5	6_38	01240	109536.7	78273.3	6	MFS transporter
6	6_39	01235	100349.3	79257.0	5	L-lactate dehydrogenase
7	3_48	02540	44267.5	18049.1	10	P159 adhesin
8	7_1	01335	38659.0	22183.2	7	2-oxoisovalerate dehydrogenase subunit beta
9	7_2	01340	37934.2	21524.6	8	pyruvate dehydrogenase (acetyl-transferring) E1 component subunit alpha
10	6_90	03425	37502.7	4180.8	39	surface lipoprotein, P65, lipolytic enzyme
11	6_96	03455	36751.8	15342.2	14	Cilium adhesin (P102 paralog)
12	6_97	03460	25312.6	6862.2	30	Cilium adhesin (P146)
13	4_84	00925	24586.3	15767.1	13	protein p97; cilium adhesin
14	3_1	02285	24475.0	15250.0	15	hypothetical protein
15	2_21	02015	22410.3	15206.3	16	PTS ascorbate transporter subunit IIC
16	13_7	02625	21366.1	12687.3	18	hypothetical protein
17	2_22	02020	21133.6	13236.9	17	phosphotriesterase
18	18_10	00760	20245.6	15971.7	12	3D-(3,5/4)-trihydroxycyclohexane-1,2-dione acylhydrolase (decyclizing)
19	4_83	01395	15054.7	8547.7	23	hypothetical protein
20	18_6	00780	14894.6	17666.8	11	methylmalonate-semialdehyde dehydrogenase

[a]Rank of genes in *in vivo* transcriptome based on expression level (RPKM)

[b]Abbreviated locus tag *M. hyopneumoniae* strain 98 (MHP_RSXXXXX)

[c]Locus tag homologous gene *M. hyopneumoniae* strain 232

[d]Average expression level in flush samples based on read numbers taking into account gene length and total amount of reads in the dataset

[e]Average expression level in culture samples based on read numbers taking into account gene length and total amount of reads in the dataset

[f]Rank of genes in *in vitro* transcriptome based on expression level (RPKM)

[g]Annotated function in strain 232

Differentially expressed genes during infection

We analyzed the distribution of sense and antisense mapping reads per gene and found on average 97.5 ± 1.2% sense mapping reads (table S3). These reads were used to analyze differentially expressed genes after filtering for genes with a read count of >100 counts per million (CPM) in two or more datasets[199]. By filtering we removed 203 genes (29% of genes in the genome). Based on FDR<0.01 and LOG2 fold change larger than two, we found 23 genes up-regulated *in vivo*, 30 genes down-regulated *in vivo* and 445 genes not differentially expressed (figure 2 and table S4). The average biological coefficient of variance calculated for all genes in the dataset using EdgeR was 0.30, meaning that on average 30% of the variation in gene abundance is the result of biological variation between biological replicates. This value for BCV is higher than expected for experiments with genetically identical organisms[198]. The PCA analysis already showed that the majority of the variation within samples types is present in the *in vivo* datasets (figure S8) and we expect that this variation is caused by varying conditions in the infected lung and variation in the genetic background of the pigs.

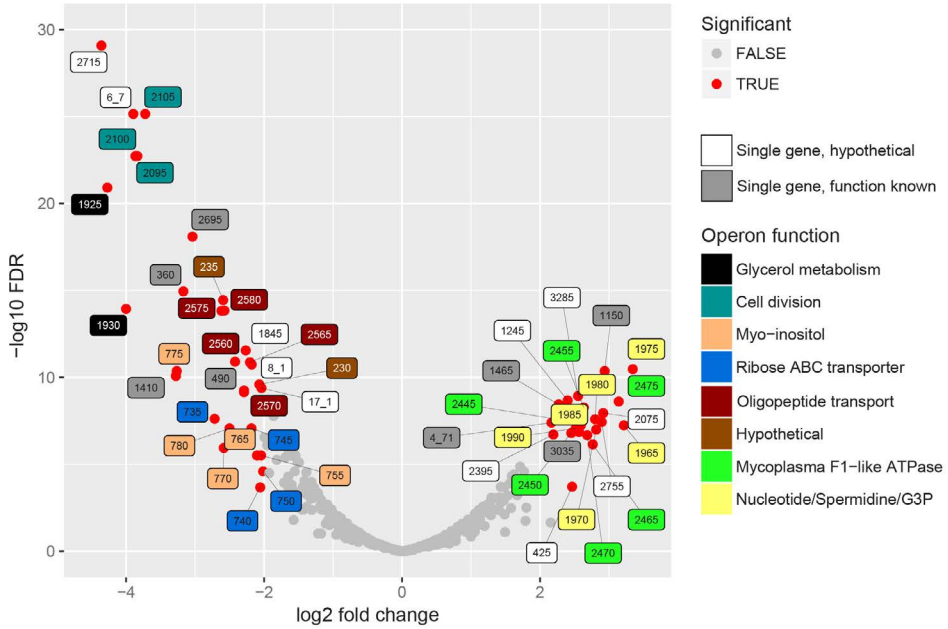


Figure 2. *In vivo* differentially expressed genes. EdgeR was used to find differentially expressed genes in the RNA seq. datasets of cultures and enriched lung flush samples. Significance in this graph is based on FDR<0.01 (dashed line) and LOG2 fold change threshold > 2. Significantly up- or down regulated genes are indicated in red, label colour shows operon structure and label names represent abbreviated strain 232 locus tags. Single genes were either not present in an operon or the operon was not differentially expressed. Table S4 shows the complete EdgeR analysis.

In vivo up-regulation of F_1-like ATPase

Genes encoding the alpha and beta subunits (MHP_RS02445 and MHP_RS02450) of F_1-like ATPase and four hypothetical genes in this operon were up-regulated *in vivo* (figure 2). Interestingly, these genes were not part of the typical operon encoding Type 1 F_1F_0ATPase but were part of the operon presumably encoding a Type 3 F_1-like ATPase[206]. Based on structural analysis[206], one of the hypothetical genes, MHP_RS02465, possibly resembles the γ-subunit of the F_1-like ATPase. Read levels for the gene that resembles the ε-subunit (MHP_RS02460) were too low and this gene was not present in the dataset. Three other hypothetical proteins (homologs of MHP_RS02455 and MHP_RS02470-MHP_RS02475) also have a possible role in formation of F_1-like ATPase, two were predicted to be present in the cell membrane, but their exact role remains to be elucidated. There is no differential expression of genes in the operon encoding the typical Type 1 F_1F_0ATPase (MHP_RS00240-MHP_RS00280). The role of the F_1-like ATPase, which originated in the Hominis group of mycoplasma species[206] in infection is unclear but might provide protection against pH stress since the ATPase is able to transport protons across the plasma membrane (figure 3). To form a functional ATPase, components from the F1-like ATPase could possibly interact with components from the standard ATPase. Alternatively, the F1-like ATPase components might have a role in translocation[206] but *M. hyopneumoniae* is not known to be motile.

Increased expression of nucleases during infection

MHP_RS01975, a putative lipoprotein expressed on the cell surface which functions as a calcium dependent exonuclease[45] was induced *in vivo* (figure 2). Exonuclease activity could be important to supply nucleotides needed for growth in the host since pathways to synthesize nucleotides *de novo* are not present in mycoplasma species. The role of nucleases in host-pathogen interaction was investigated in multiple mycoplasma species: nucleases from *M. hyorhinis* caused apoptotic symptoms in a human pancreatic adenocarcinoma cell line[207], endonuclease P40 expressed by *M. penetrans* had a cytotoxic effect on lymphocytic cell lines[208] and MAG_5040 of *M. agalactiae*, which shows sequence similarity with MHP_RS01975, was shown to degrade the DNA component of neutrophil extracellular traps[209]. We also found up-regulation of the A subunit of excinuclease ABC (uvrABC) involved in DNA repair (MHP_RS01465) and of DNA polymerase I (MHP_RS03035). This could indicate that DNA damage is occurring more frequently under *in vivo* conditions for instance due to the oxidative stress induced by neutrophils. We also found up-regulation of MHP_RS01150, a ribosome small subunit-dependent GTPase, but a possible role during infection for this protein is unclear.

In vivo up-regulation of genes related to glycerol-3-phosphate transport and spermidine transport

Other genes present in the MHP_RS01975 operon were also up-regulated. MHP_RS01985 and MHP_RS01990 were annotated as putative glycerol-3-phosphate transport system permeases (ugpA and ugpE) (figure 3). Glycerol-3-phosphate could be used as an alternative carbon source (figure 3) or for cardiolipin biosynthesis. Also present in this operon is MHP_RS01980 (*potA*) which is a potential spermidine/putrescine transporter. Finally, two up-regulated genes in the operon were putative lipoproteins: MHP_RS01965 and MHP_RS01970, with an unknown function and no further functional information could be derived from the protein domain composition of these proteins.

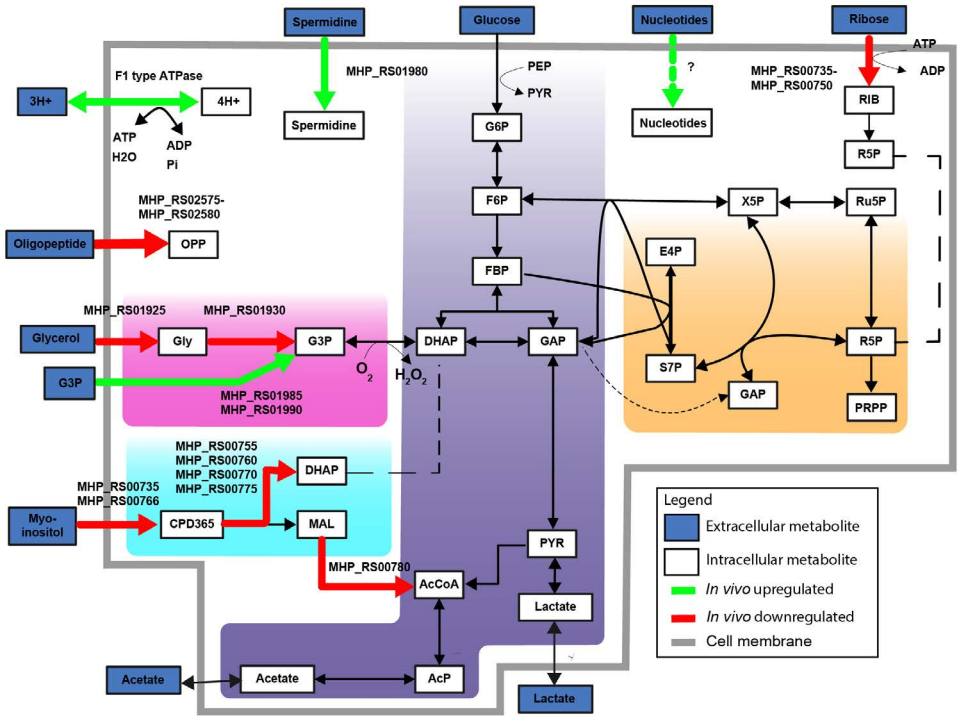


Figure 3. F1 type ATPase and glycerol-3-phosphate transport upregulated *in vivo*. Differentially expressed metabolic genes are shown, up- and downregulated genes are shown based on FDR<0.01 and LOG2 fold change threshold > 2. Upregulated metabolic genes *in vivo* (green). Downregulated metabolic genes *in vivo* (red). Highlighted simplified pathways are glycolysis (blue), pentose phosphate pathway (orange), glycerol (pink) and myo-inositol metabolism (turquoise). Abbreviations are explained in table S11 of chapter 3.

P102 cilium adhesin down-regulated during infection

Surprisingly, the gene encoding the cilium adhesin P102 (17_1, MHP_RS00920) was repressed during infection. Adhesion to the cilia is important for *M. hyopneumoniae* to reside on the respiratory epithelium. We also identified down-regulation of P97 cilium

adhesin but with a <2LOG2 fold change (table S4). As the bacterial RNA was isolated from bacteria already adhered to the ciliated epithelium, this could indicate a down-regulation upon contact with cilia. Alternatively, transcription could be up-regulated in the *in vitro* culture when cells were grown in suspension. Another possible explanation is that by flushing lung lobes we selected for bacteria that were not attached to the cilia resulting in isolation of bacteria with lowered expression of P102.

Genes related to cell division were down-regulated during infection

The operon with genes related to cell division (MraZ, RsmH and ftsZ) was down-regulated *in vivo* (figure 2). This could indicate a lower growth rate *in vivo* when compared to *in vitro* growth in complex medium. Alternatively, part of the bacterial population isolated by flushing could be in the stationary phase (or even starvation phase) instead of growth phase as these genes were found to be down-regulated in *M. pneumoniae* cultures when entering stationary phase[55]. We also found down-regulation of MHP_RS00360, chaperone DnaK (70 kDa heat-shock protein) and chaperone protein ClpB (MHP_RS01410). DnaK binds to denatured proteins during heat-shock and ClpB in general binds misfolded and aggregated proteins. Our culture conditions were not at high temperature, possibly these proteins play a role during normal growth and are down-regulated in relation to the growth rate. Genes MHP_RS00490, "MurR/RpiR family transcriptional regulator" and a Type IV secretory system conjugative DNA transfer protein, "ICEF-IIA" (MHP_RS02695) were also found down-regulated but the role of these genes in *M. hyopneumoniae* during infection is not understood.

Genes related to alternative carbon metabolism were down-regulated in vivo

There was down-regulation under *in vivo* conditions of the glycerol uptake facilitator and the glycerol kinase (figure 3, homologs of MHP_RS01925 and MHP_RS01930). Glycerol is a possible energy source and is used for production of cardiolipin. Glycerol consumption by mycoplasma species has been associated with virulence because hydrogen peroxide is released when glycerol-3-phosphate is converted to di-hydroxyacetone phosphate via the glycerol oxidase[36]. There was also down-regulation *in vivo* of the operons related to myo-inositol and ribose transport and catabolism (figure 3). Another down-regulated operon is involved in oligopeptide transport. This could indicate that oligopeptides, which are present in a high concentration in the complex growth medium, are absent or present at very low levels in the natural niche. Transport of single amino acids could alternatively provide *M. hyopneumoniae* with amino acids needed for protein synthesis.

ncRNAs were differentially expressed during infection

The role of non-coding RNAs in *M. hyopneumoniae* is not known. Given the high AT-content of the genome it was expected that most antisense RNAs found in our datasets were the result of spurious transcription (chapter 6). However, some might have a regulatory role

and these could play a role during infection. Although the amount of antisense reads in our datasets was low (table S3), we analyzed which ncRNAs were differentially expressed using the same methodology as used to find differentially expressed genes. Based on a best-blast hit (e-value cut-off $<1x10^{-6}$ and ranked according to bit-scores correlated to query length) with the strain 98 genome, we identified 606 out of 629 ncRNAs annotated in strain 232 (chapter 6). To the annotated ncRNAs on average 18.0 ± 1.5% of the antisense mapping reads could be mapped. By filtering for ncRNAs with a low read count we removed 105 ncRNAs from our analysis. Of the remaining ncRNAs we found 28 up-regulated *in vivo*, 66 down-regulated *in vivo* and 407 not differentially expressed (figure S9) based on FDR<0.01.

Discussion

In this study we sequenced the genome and determined the *in vivo* transcriptional landscape for *M. hyopneumoniae* strain 98 using RNA sequencing after applying a novel method to isolate bacterial RNA which allowed us to reach sufficient sequencing depth. We compared the *in vivo* transcriptome to culture grown cells and found up-regulated genes *in vivo* with functions related to F_1-ATPase, nucleotide metabolism and glycerol-3-phosphate transport while down-regulated genes were related to cilium adhesion, cell division, glycerol metabolism and alternative carbon metabolism (myo-inositol and ribose). We showed that the enrichment step for bacterial RNA in our protocol did not introduce bias for determination of the *in vivo* transcriptome. Functional analysis of the differentially expressed genes could further increase our understanding of the infectious mechanisms used by *M. hyopneumoniae*.

Although RNA sequencing has become the method of choice for studying bacterial transcriptomes, the technical limitations of this method should be kept in mind. The amount of input material for the *in vivo* samples in this study was at the lower limit accepted for library preparation and this could have caused measurement variation between the different sample types. Although we improved the experimental protocol for *in vivo* sampling, still on average 96% of the reads obtained could not be mapped to the bacterial genome and insufficient reads were obtained for 29% of the genes in the bacterial genome. Further increasing the sequencing depth or further optimization of the sampling method should allow us to analyze the expression level changes of some of these genes. Also, the high concentration of background RNA could have caused a bias in our experiment because the amount of reads mapping to a gene depends on the gene length, expression level and the composition of the RNA sample[200]. We normalized for sequencing depth but the high sequencing capacity needed for the background RNA could have caused underrepresentation of the *M. hyopneumoniae* reads in the *in vivo* datasets.

This is the first time the *in vivo* transcriptome of *M. hyopneumoniae* was analyzed using RNA sequencing. We compared our results to the microarray analysis performed by

Madsen *et al.*[189] We found no correlation (Spearman's coefficient = -0.14, figure S10) between fold changes for 162 differentially expressed genes reported in the microarray study and fold changes for the same genes in our study. There could be various explanations for the lack of correlation between the two studies: (1) different *M. hyopneumoniae* strains were used, (2) pigs had a different genetic background and age, (3) different methods were used to isolate bacteria and bacterial RNA and (4) presence of host RNA could have caused non-specific binding in the microarray study. It was recently shown that differentially expressed genes at low transcript levels could be identified with RNAseq and validated with qPCR while differentially expressed genes identified with microarrays could not be validated with qPCR when transcript levels were low[210]. Since the lung flush samples contained a high quantity of background RNA and low bacterial transcript levels, RNAseq seems a more suitable method to determine bacterial transcript levels in this sample type.

In this study we found differentially expressed non-coding RNAs. The function of these ncRNAs was not known but could be further elucidated by studying their expression levels under various culture conditions or their essentiality by random mutagenesis. In a recent study in *M. pneumoniae* it was reported that the majority of ncRNAs were likely formed as a result of spurious transcription (chapter 6) at a minimal expense of cellular energy. However, some ncRNA's were found to be essential in *M. pneumoniae* and a small amount of ncRNAs were actually small open reading frames and could be translated into proteins[100]. Regulatory ncRNAs could act in a *cis*-manner (when the ncRNA sequence overlaps the gene sequence it regulates), in a *trans*-manner (when the ncRNA and target gene sequence do not overlap) or by binding to proteins[211]. Regulation by ncRNAs is often mediated by Hfq, for instance in the model pathogen *Salmonella enterica* serovar Typhimurium multiple ncRNAs bind Hfq and play a role in expression of the outer membrane proteins which are important during infection[212]. However, Hfq is not found in *M. hyopneumoniae* and it is not known if the function of Hfq could be performed by an alternative protein. Regulation without Hfq mediation was also described in *Salmonella*, for instance AmgR, a *cis*-encoded 1.2 kb ncRNA binds the *mgtC* portion of the *mgtCBR* mRNA and regulates response to low magnesium concentrations[213]. It would be highly interesting to gain further insight on the role of ncRNAs in *M. hyopneumoniae* during infection.

The bacterial transcriptional landscapes presented provide insight in the gene expression levels of the whole bacterial population that was isolated. Within the lung, but also in *in vitro* systems, population differences will exist due to changes in the micro-environment (e.g. nutrient availability, pH, temperature or varying interactions with the host). We expect that further development of single cell RNA sequencing techniques will enable us to elucidate these population differences in the near future. Sequencing a single bacterial cell isolated from the lung will also remove the (host) background RNA and decrease the amount of sequencing capacity needed. It would also be interesting to study

the bacterial transcriptome in different pig breeds or pigs of different ages since in our study all pigs were of the same breed and age.

The exact role in the infection process of the differentially expressed genes identified in this study should be further investigated. This can be done in cell culture systems infected with *M. hyopneumoniae* in which the conditions in the lung are mimicked and transcriptional landscapes could be compared with the transcriptional landscape in the lung. Also, the importance of some of the differentially expressed metabolic pathways could be verified by changing culture conditions in such *in vitro* systems, for instance by changing the availability of myo-inositol or ribose. For development of novel or more effective vaccines, the up-regulated genes found in this study could be considered for use in subunit vaccines or live-attenuated mutant strains, for use in live vaccines, could be created by gene-knockout of *in vivo* up-regulated genes. These are multiple ways to use the novel insight into the *M. hyopneumoniae* transcriptional landscape to further investigate pathogenic mechanisms.

Nomenclature

Metabolite abbreviations are explained in table S11 of chapter 3.

Acknowledgements

Maria Suarez-Diez for help with analysis of data quality using PCA. We thank our farmers, animal care-takers and veterinarian for excellent care and handling of our pigs and thank our colleagues from Swine R&D for supporting this study.

Supplementary figures

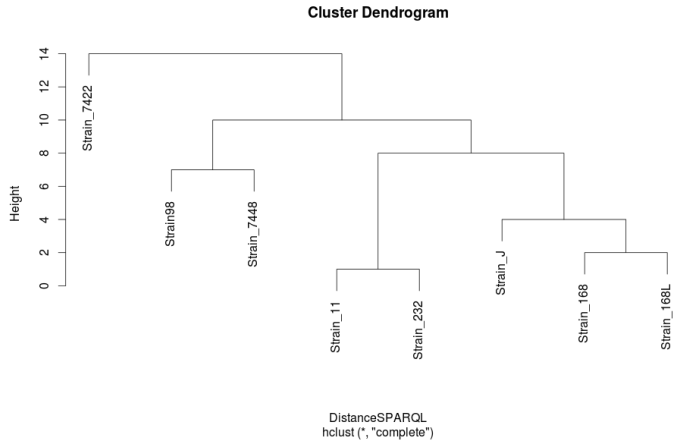


Figure S1: functional clustering based on Manhattan distance calculated from the presence/absence matrix of the *M. hyopneumoniae* accessory domainome of multiple strains

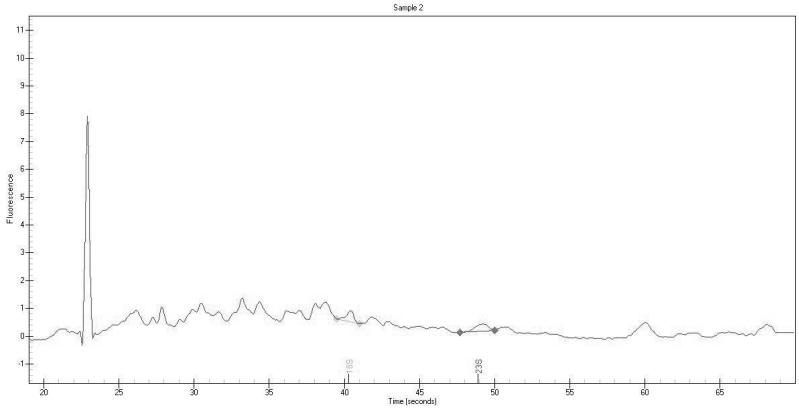


Figure S2: Experion profile sample F1-E

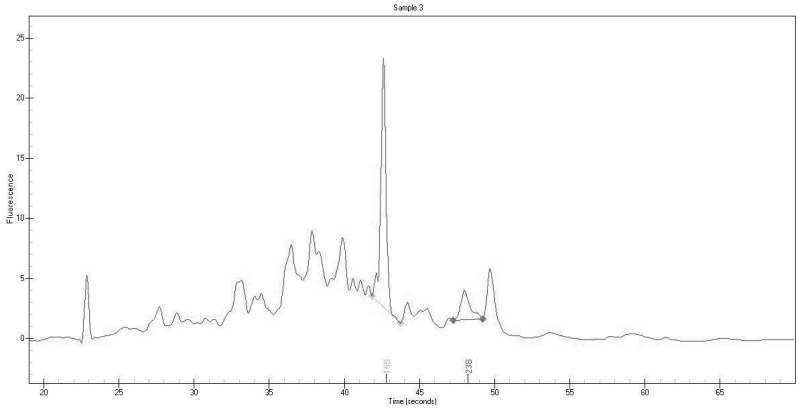


Figure S3: Experion profile sample F1-NE

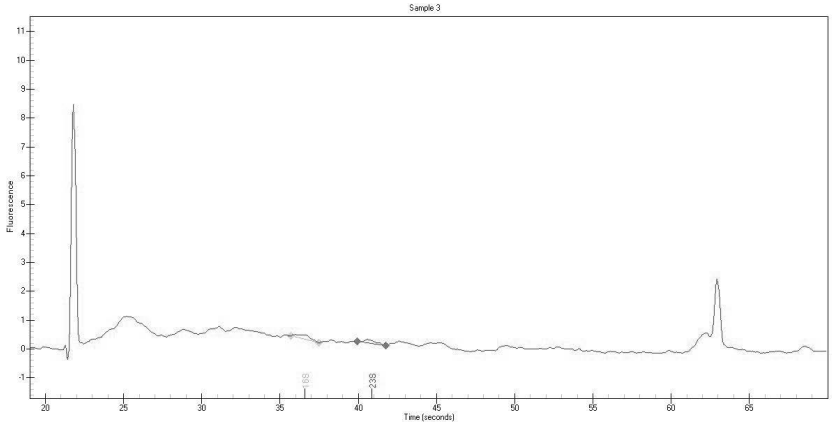


Figure S4: Experion profile sample F2-E

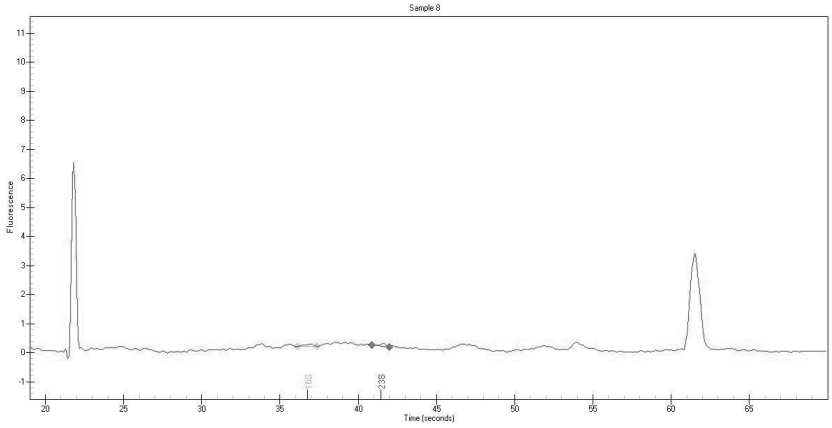


Figure S5: Experion profile sample F3-E

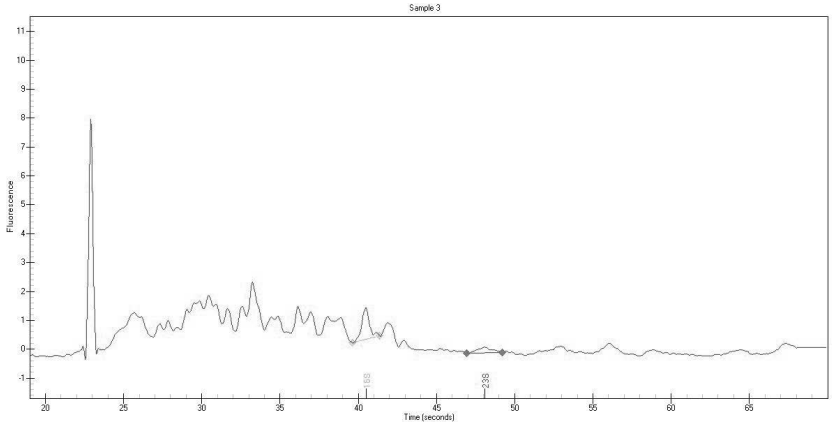


Figure S6: Experion profile sample F4-E

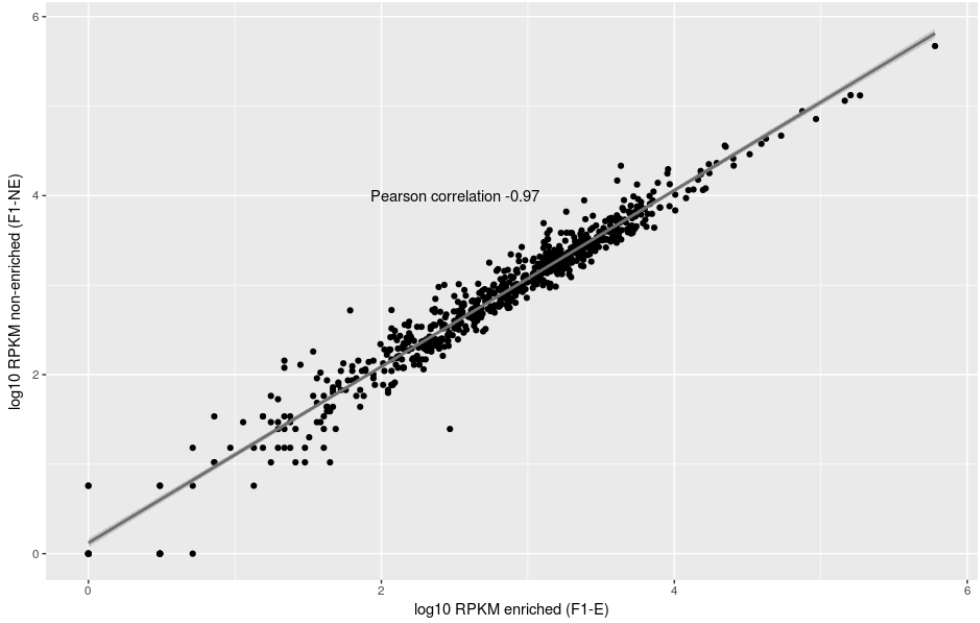


Figure S7: correlation analysis between expression levels of genes (log 10 RPKM) in an enriched flush sample (F1-E) and a non-enriched flush sample (F1_NE) sample, the same flush sample was used as input for RNA purification and sequencing, all steps in the sample treatment were followed.

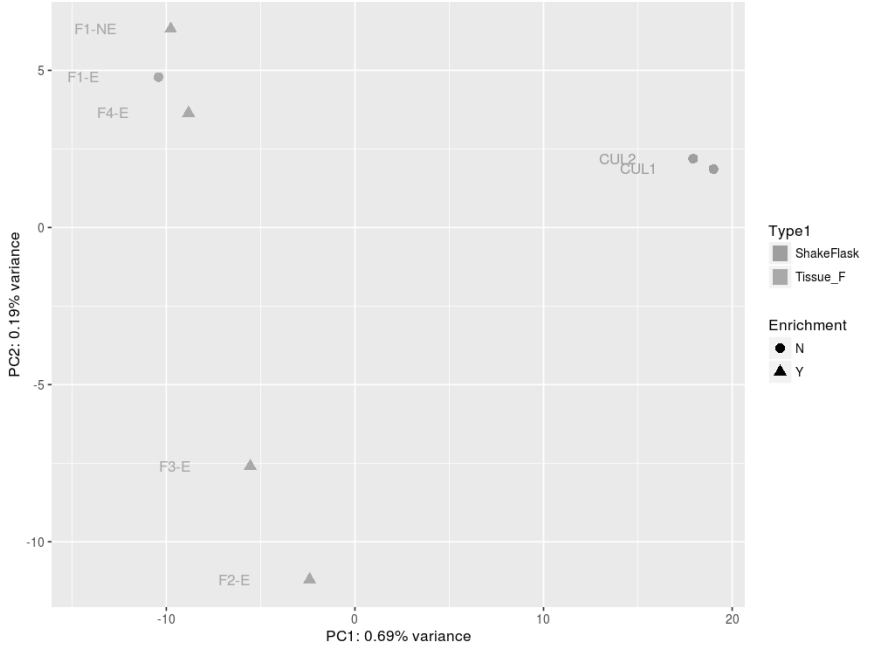


Figure S8: PCA analysis (selected flush samples, tissue and culture samples).

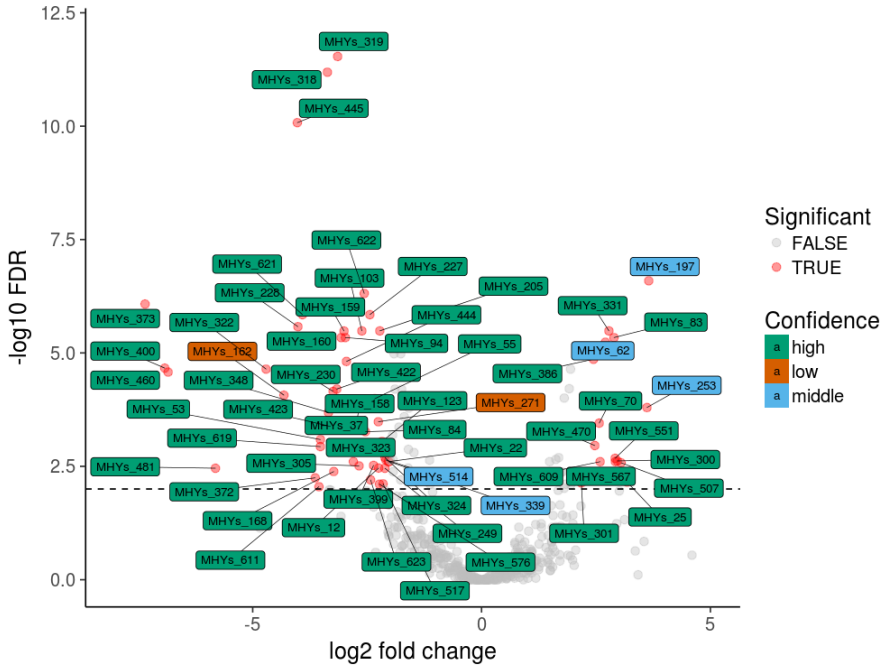


Figure S9: Volcano plot showing differentially expressed non-coding RNAs. Significance was determined based on fold change (>2 log2 up- or downregulated) and FDR<0.01. Confidence score is indicate and is based on blast data output (bit-score divided by query length: <1=low, 1-1.5=middle and >1.5=high)

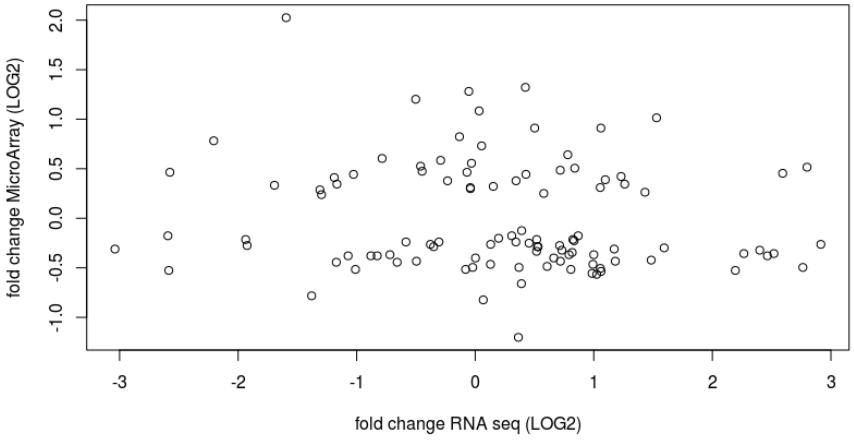


Figure S10: correlation between log2 fold changes reported for DE genes by Madsen *et al*[189] and the same genes in our study.

Supplementary tables

Table S1. Genome properties of *Mycoplasma hyopneumoniae* determined using SAPP

Strain	Total amount of proteins[a]	Proteins with domains[b]	Total unique domains[c]
Strain_11	681	466	855
Strain_232	681	468	856
Strain_J	692	479	859
Strain_7448	690	476	862
Strain_168	698	470	857
Strain_168L	700	471	857
Strain98	694	471	857
Strain_7422	703	477	857

[a]Amount of proteins based on prodigal gene calling

[b]Number of proteins which contain at least one protein domain

[c]Number of unique domains annotated by InterProScan

Table S2 is provided online at http://edepot.wur.nl/425152, DOI: 10.18174/425152.

Table S3. Determination of percentage sense mapping reads in vivo and in vitro

Sample	Sample type	Total reads	Total mapping	Mapping genes	Mapping sense	Mapping antisense	% Sense mapping
F1-E	*In vivo*	37327323	1372521	1334103	1276640	57463	95.7
F2-E	*In vivo*	94501247	2878131	2225857	2182660	43197	98.1
F3-E	*In vivo*	76920873	1853172	1732281	1689410	42871	97.5
F4-E	*In vivo*	38501950	2616927	2561616	2469070	92547	96.4
CUL1	*In vitro*	24327141	24133581	22992127	22642996	349131	98.5
CUL2	*In vitro*	31860823	31585953	29317237	28894924	422313	98.6

Table S4 is provided online at http://edepot.wur.nl/425152, DOI: 10.18174/425152.

Chapter 6

Bacterial antisense RNAs are mainly the product of transcriptional noise

Verónica Lloréns-Rico, Jaime Cano, Tjerko Kamminga, Rosario Gil, Amparo Latorre, Wei-Hua Chen, Peer Bork, John I. Glass, Luis Serrano and Maria Lluch-Senar

Published in Science Advances

Abstract

Cis-encoded antisense RNAs (asRNAs) are widespread along bacterial transcriptomes. However, the role of the vast majority of these RNAs remains unknown, and there is an ongoing discussion as to what extent these transcripts are the result of transcriptional noise. We show, by comparative transcriptomics of 20 bacterial species and one chloroplast, that the number of asRNAs is exponentially dependent on the genomic AT content, and that expression of asRNA at low levels exerts little impact in terms of energy consumption. A transcription model simulating mRNA and asRNA production indicates that the asRNA regulatory effect is only observed above certain expression thresholds, substantially higher than physiological transcript levels. These predictions were verified experimentally by overexpressing 9 different asRNAs in *M. pneumoniae*. Our results suggest that most of the antisense transcripts found in bacteria are the consequence of transcriptional noise, arising at spurious promoters throughout the genome.

Introduction

The catalog of bacterial-encoded RNAs has recently undergone a vast expansion. The canonical mRNAs and known non-coding RNAs (rRNAs, tRNAs, tmRNA and others) are now accompanied by a handful of new transcript categories. Small, non-protein-coding RNAs or sRNAs are one of these new categories. The numbers of initially reported sRNAs ranged from dozens to hundreds in different species[214,215]. These include cis-encoded sRNAs, which overlap functionally defined genes, either in sense or antisense (thus named asRNAs), and trans-encoded sRNAs, which are separated from their target genes. These sRNAs span a wide range of lengths: from dozens to a few thousand base pairs[215]. However, recent improvements in techniques for analysis of transcription have revealed that non-coding transcription in prokaryotes is pervasive through the genome[216–218]. Still, only few sRNAs have been functionally characterized[219–221], most of which correspond to the category of trans-encoded sRNAs. Examples of these are the ones associated with bacterial virulence[211,222,223]. The most common mechanism of action of sRNAs is via complementary base pairing with coding sequences (Fig. S1a). RNA duplex formation between sRNA and mRNA can change mRNA stability, inducing degradation or stabilization of the duplex. This duplex may as well induce or repress mRNA translation by affecting the ribosome binding site[215,224]. Another asRNA regulatory mechanism is transcriptional interference, occurring if two RNA polymerases transcribing in convergent directions collide[225]. Other types of RNA having a regulatory role by 'non-standard' mechanisms should not be discarded. For instance, if there was a Dicer-like mechanism in bacteria as it occurs in eukaryotes [226], low abundant RNAs could exert a strong influence on complementary, more abundant, mRNAs. In this respect we have the CRISPR/Cas system, where crRNAs, even if not abundant, target the enzyme against foreign DNA[227] and/or RNA sequences[228].

There is an ongoing discussion both in eukaryotes and prokaryotes as to what extent this plethora of ncRNAs provides a crucial layer of transcriptional and translational regulation, or if a large part of them are the result of transcriptional noise, arising from spurious promoters[229,230]. Bacterial promoters are characterized by low information content, and their major landmark is the Pribnow motif that has the consensus sequence TANAAT[231]. Other features include: i) the -35 box, although this has been shown not to be essential (especially in *Firmicutes*) and can be replaced by other elements[232]; and ii) low melting energies, which ultimately depend on the AT composition of the promoter region. Such low information content implies that promoters could easily arise by random mutations in bacterial genomes, especially given the presumptive bias towards G/C nucleotides mutating to A/T[233]. If sRNAs are the product of transcriptional noise due to spurious TANAAT boxes, we predict that the number of sRNAs in bacteria will strongly correlate to the AT content of their genomes in an exponential manner. (Fig. S2a). Random production of asRNA from these spurious Pribnow boxes at low levels could, by the

stochastic nature of transcription, and the short half-life of RNAs in bacteria, not affect the levels of the sense mRNA (Fig. S1b).

Materials and methods

Bacterial strains and growth conditions

M. hyopneumoniae:
Culture samples from *M. hyopneumoniae* were obtained from batch fermentation in exponential growth. 50 ml culture was centrifuged for 3 minutes at 9000Gs in a cooled centrifuge (2-8°C). Supernatant was removed and the cell pellet was stabilized using RNA later (Ambion). Stabilized cell pellets were stored at 2-8°C until RNA extraction.

B. aphidicola:
Cedar aphids were collected from a population maintained in the facilities of the Institut Cavanilles de Biodiversitat i Biologia Evolutiva at the University of Valencia (ICBiBE, Paterna, Valencia, Spain. 39° 30' 57.5598" N, 0° 25' 20.0892" W[234]).

M. pneumoniae:
M. pneumoniae was grown in 50 mL of modified Hayflick medium supplemented with glucose at 37ºC as previously described[235]. To select mycoplasma cells expressing the ncRNAs, medium was supplemented with 2μg ml^{-}1 tetracycline.

M. mycoides:
M. mycoides JCVI syn1.0[141] was grown in 50 ml of SP4 medium containing 17% fetal bovine serum at 37ºC and harvested during to mid-log phase as previously described[236].

RNA extraction

M. hyopneumoniae:
RNA was extracted from a bacterial pellet stabilized with RNAlater (Ambion) using the Quick-RNA MiniPrep (Zymo Research) following manufacturer's protocol.

B. aphidicola:
The bacteriomes of 200 adult wingless parthenogenetic insects were dissected under a Wild Heebrugg Plan 1X microscope and preserved on RNAlater (Ambion) at -80ºC until its use. The bacteriome sample was defrost and washed with PBS (137 mM NaCl, 2.7 mM KCl, 10 mM Na2HPO4, 2 mM KH2PO4, pH7.2), and total RNA was purified using the TRI Reagent Solution Kit (Ambion).

M. pneumoniae:
After growing *M. pneumoniae* strains for 6h at 37°C, cells were washed twice with PBS and lysed with 700 µl of Qiazol buffer. Then, samples were lysed with 700 µl of Qiazol buffer. RNA extractions were performed by using the miRNeasy mini Kit (Qiagen) following the instructions of the manufacturer.

M. mycoides:
Cells were centrifuged from culture medium, and washed twice in hepes buffered saline containing 20% sucrose. Cell pellets were stabilized with RNAprotect (Qiagen) until extraction with UltraClean RNA isolation kits (MoBio).

Library preparation and RNA sequencing

M. hyopneumoniae:
Ribosomal RNA was removed using the Ribo-Zero kit (Epicentre). rRNA-depleted RNA was fragmented with an average length of 100-200bps and converted to double-stranded cDNA. Library preparation was done using a protocol based on the "dUTP method", to generate strand-specific mRNA-seq libraries including barcoding[237,238]. The Illumina stranded TruSeq RNA-seq library preparation kit was used. Sequencing of the library was done using the Illumina HiSeq: single-end reads, one lane, 50 cycles, two samples per lane. The sequencing data produced was processed removing low-quality sequence reads. Additionally, the sequence data in FastQ format was additionaly filtered and trimmed based on Phred quality scores.

B. aphidicola:
The samples were mRNA-enriched using the MicrobExpress Kit (Ambion) and the MAgentig Kit (Epicentre) to remove rRNA of bacterial and eukaryote origin, respectively, following the manufacturer's protocols. Library preparation was done with the SOLID Total RNA-Seq Kit (Life Technologies), and sequencing was performed with an ECC Module on a 5500 XL Genetic Analyzer (Life Technologies) at the sequencing facility of the University of Valencia.

M. pneumoniae:
Libraries for RNA-seq were prepared following directional RNA-seq library preparation and sequencing as previously described[100].

M. mycoides:
cDNA libraries were constructed with ScriptSeq Complete Gold kits (Epicentre Biosciences) and were sequenced on an Illumina HiSeq instrument.

6.1.1 Data analysis

Reads from all RNA-seq experiments detailed above were mapped to their corresponding reference genomes using MAQ[239]. All reads were treated as single-end reads. For paired-end sequencing reads, only fragment 1 was considered. After mapping, only reads mapping to a unique position in the reference genome were used. Pileups were obtained using a custom-designed software and visualized on the Integrative Genomics Viewer (IGV[240]) for manual annotation of sRNAs. An example of this manual annotation using IGV can be found in Fig. S2.

In order to filter out noise and define the regions corresponding to sRNAs, we first determined the expression levels of all ORFs in the different bacteria analyzed. Expression levels were calculated using a custom-made script to determine the CPKM (Counts Per Kilobase per Million Counts) values, a measure that is similar to RPKM for single-end reads. For each genome, we used the expression values of genes with known function. The lower 0.05 quantile of this distribution was chosen as a threshold to determine expression of new non-annotated features. Trans-encoded sRNAs and asRNAs above this threshold were manually identified and annotated.

Regarding the published data from other bacterial species, whenever the sRNA annotation was available, we mapped the non-coding transcripts to the genome to determine how many of them overlapped a gene in antisense, and how many corresponded to trans-encoded sRNAs. To do so, we used the reference annotations from the NCBI to define ORFs. Partial or total overlap was considered, and for bacteria with more than one replicon, only features in the largest replicon were considered. In some cases, the numbers of asRNAs and trans-encoded sRNAs differ from the numbers reported in the different publications. This is due to the usage of different annotation versions, the inclusion or not of UTRs, and the consideration of all the replicons in the different publications.

Calculation of the energy cost of non-coding RNA transcription

To determine the energy cost of transcribing non-coding RNAs, we estimated the relative cost compared to the transcription of mRNAs, rRNAs and tRNAs. The cost of transcription was assumed to be proportional to the average length of the RNAs multiplied by their transcription rates. In *M. pneumoniae*, there are 738 ORFs, 3 rRNAs and 37 tRNAs, The average length of each of these groups is 981.38, 1516 and 77.91 base pairs respectively. Transcription rates for each group were estimated from an equilibrium situation, as follows:

$$\frac{dm}{dt} = \alpha_m - k_m[m]$$

Where [m] is the mRNA concentration, α_m is the transcription rate, and k_m the decay rate. In equilibrium:

$$\frac{dm}{dt} = 0$$

$$\alpha_m = k_m[m]$$

RNA concentrations in exponential growth were estimated using the copy numbers previously reported[241], and extrapolating to all RNAs in the cell according to experimental RNA-seq data. RNA decay rates were experimentally determined using novobiocin, a DNA gyrase inhibitor, which releases the RNA polymerase from the chromosome[56]. After the treatment with this inhibitor, RNA from the cells was extracted at different timepoints and RNA concentrations were determined[56]. RNA decay in the cell population was thus modeled following an exponential decay, as follows:

$$\frac{dm}{dt} = \alpha_m - k_m[m]$$

After the treatment:

$$\alpha_m = 0$$

$$\frac{dm}{dt} = -k_m[m]$$

Solving this we fitted our experimental data to the following exponential decay:

$$[m]_t = [m]_0 \cdot e^{-k_m t}$$

And obtained the degradation rate values, k_m. Averages for mRNAs and asRNAs not overlapping other transcripts were used, to ensure no other factors participate in the degradation. However, we compared the transcription and decay rates determined experimentally between genes overlapped by asRNAs and genes not overlapped by any other transcript, and strikingly we found no statistically significant differences in transcription rates (p-val=0.29, Mann-Whitney U test) or decay rates (p-val=0.053, Mann-Whitney U test).

Transcription rates were estimated to be, on average, 0.016 molecules/min, 0.966 molecules/min and 0.061 molecules/min for mRNAs, rRNAs and tRNAs respectively. An estimate of the energy that the cell spends in transcribing these molecules can be obtained by multiplying their number by their length and their transcription rate. Multiplying these values we got an estimate of 16157.37 (a.u). Following the same logic for sRNAs, in *M. pneumoniae* there are 251 sRNAs, with an average length of 270.597bps and a transcription

rate of 0.007 molecules/min. Multiplying these values we got an estimate of 475.43 (a.u), equivalent to 2.94% of the energy spent in transcribing mRNAs, tRNAs and rRNAs together.

Previous studies report that the energy spent in transcribing total RNA (referring to mRNAs, tRNAs and rRNAs), in terms of number of ATPs required, is ~5000 units of ATP per second per cell[67]. This implies that, according to our calculations, the number of ATPs required for sRNA transcription would be ~147 units per cell per second. This number, compared with the total ATP produced by the cell (~60000 units per second in mid-exponential growth[67]), results in only 0.24% of the cell's generated energy.

A similar calculation was performed in *E. coli*. The genome of *E. coli* codes for 4067 genes, with an average length of 907.09bps, and 1005 asRNAs[242]. Because of the lack of a complete annotation of these asRNAs, we used the average length of the sRNAs in *M. pneumoniae*. The approximate transcription rate used was ~0.0602-0.602 molecules/min[243]. Assuming both genes and asRNAs are transcribed at 0.602 molecules/min, the energy E. coli spends in antisense transcription is 6.7% of that spent in sense transcription. If we consider that transcription of asRNAs occurs at a lower rate of 0.0602 molecules/min, this percentage decreases to 0.67% of energy spent in antisense transcription.

Mathematical modeling of the effect of the asRNAs

Three putative effects of the asRNAs were considered: in case 1, the binding of the asRNA to the corresponding mRNA induces degradation of the duplex. In case 2, the binding of the asRNA to the mRNA induces degradation of the mRNA, but not of the asRNA. In case 3, the mRNA and the asRNA bind reversibly to form a stable duplex, preventing translation of the mRNA. In the three cases, binding to the ribosome protects the mRNA from the effect of the asRNA. The three cases were modeled as follows:

Case 1:

$$\frac{dm}{dt} = \alpha_m + \alpha_p[m^{rib}] - \beta[m][rib] - k_m[m] - k_{on}[m][s]$$

$$\frac{ds}{dt} = \alpha_s - k_s[s] - \gamma[m][s]$$

$$\frac{drib}{dt} = \alpha_p[m^{rib}] - \beta[m][rib]$$

$$\frac{dm^{rib}}{dt} = \beta[m][rib] - \alpha_p[m^{rib}]$$

$$\frac{dp}{dt} = \alpha_p[m^{rib}] - k_p[p]$$

Case 2:

$$\frac{dm}{dt} = \alpha_m + \alpha_p[m^{rib}] - \beta[m][rib] - k_m[m] - k_{on}[m][s]$$

$$\frac{ds}{dt} = \alpha_s - k_s[s]$$
$$\frac{drib}{dt} = \alpha_p[m^{rib}] - \beta[m][rib]$$
$$\frac{dm^{rib}}{dt} = \beta[m][rib] - \alpha_p[m^{rib}]$$
$$\frac{dp}{dt} = \alpha_p[m^{rib}] - k_p[p]$$

Case 3:

$$\frac{dm}{dt} = \alpha_m + \alpha_p[m^{rib}] - \beta[m][rib] - k_m[m] - k_{on}[m][s] + k_{off}[dup]$$
$$\frac{ds}{dt} = \alpha_s - k_s[s] - k_{on}[m][s] + k_{off}[dup]$$
$$\frac{drib}{dt} = \alpha_p[m^{rib}] - \beta[m][rib]$$
$$\frac{dm^{rib}}{dt} = \beta[m][rib] - \alpha_p[m^{rib}]$$
$$\frac{dp}{dt} = \alpha_p[m^{rib}] - k_p[p]$$
$$\frac{ddup}{dt} = k_{on}[m][s] - k_{off}[dup] - k_{dup}[dup]$$

In the equations above, *m* stands for the mRNA concentration, *s* for the asRNA concentration and *p* for the protein concentration. *rib* stands for the ribosome concentration, *dup* for the duplex concentration and m^{rib} for the mRNA-ribosome complex concentration. The values of all the parameters of the model are summarized in Table S8, and the majority of them were determined specifically for *M. pneumoniae*. RNA decay rates were determined experimentally (see previous section[56]). Using the experimental decay rates and assuming an equilibrium situation (see previous section) we determined experimental transcription rates for all RNAs in the *M. pneumoniae*. RNA concentrations used in the calculations had been previously reported[241]. All simulations were run for a time of 1000 minutes using Matlab. An SBML version of each of the models was generated using COPASI[244] and has been submitted to the BioModels database[129].

DNA manipulations and transformation of *M. pneumoniae*.

Different ncRNAs encoded by *M. pneumoniae* genome were PCR amplified by using primers described in Table S9. All 5'-primers included sequence of the constitutive promoter (P438) that drove the overexpression of ncRNAs. PCR fragments were inserted into the pMTnTetM438 minitransposon[245] by Gibson Assembly. Transformation of *M.*

pneumoniae M129 strain was performed as previously described[235] and clones were selected by supplementing the medium with 2 μg ml^{-1} tetracycline at 37°C in 5% CO_2.

Proteomics data obtaining and analysis

M. pneumoniae strain M129 was grown for 6 h at 37°C. The medium was then removed, and cells were washed twice with PBS. Total protein extract was obtained by breaking the cells with 200 μl of lysis buffer (4% SDS, 0.1M DTT and 0.1M Hepes). Total protein extracts of two biological replicates were analyzed by MS.

Each fraction (with amounts ranging from 20 to 486 μg) was trypsin-digested. Briefly, samples were dissolved in 6 M urea, reduced with DTT (10 mM, 37 ºC, 60 min), and alkylated with iodoacetamide (20 mM, 25ºC, 30 min). Samples were diluted 10-fold with 0.2 M NH_4HCO_3 before being digested at 37 ºC overnight with trypsin (with a protein:enzyme ratio of 10:1). Peptides generated in the digestion were desalted, evaporated to dryness, and dissolved in 300 μl of 0.1% formic acid. An aliquot of 2.5 μl of each fraction (amounts ranging from 0.17 to 4 μg) was run on an LTQ-Orbitrap Velos (ThermoFisher) fitted with a nanospray source (Thermofisher) after a nanoLC separation in an EasyLC system (Proxeon). Peptides were separated in a reverse phase column, 75 μ m x 150 mm (Nikkyo Technos Co., Ltd.) with a gradient of 5 to 35% acetonitrile in 0.1% formic acid for 60 min at a flow of 0.3 mL/min. The Orbitrap Velos was operated in positive ion mode with nanospray voltage set at 2.2 kV and source temperature at 325 °C. The instrument was externally calibrated using Ultramark 1621 for the FT mass analyzer and the background polysiloxane ion signal at m/z 445.120025 was used as lock mass. The instrument was operated in data-dependent acquisition (DDA) mode and full-MS scans were acquired in all experiments over a mass range of m/z 350-2000 with detection in the Orbitrap mass analyzer set at a resolution setting of 60.000. Fragment ion spectra produced via collision-induced dissociation (CID) were acquired in the ion trap mass analyzer. In each cycle of data-dependent analysis, the top twenty most intense ions with multiple charged ions above a threshold ion count of 5000 were selected for fragmentation at normalized collision energy of 35% following each survey scan. All data were acquired with Xcalibur 2.1 software. Total extract (20 μg) was also digested and desalted, and 1 μg of the resulting peptides were analyzed on an Orbitrap Velos Pro in the same conditions as the fractions but with a longer gradient (120 min).

Protein identification was performed by Proteome Discoverer software v.1.3 (ThermoFisher) using MASCOT v2.4.01 (Matrix Science) as search engine[246]. MS/MS spectra were searched against a HomoConTrans19 database comprising all putative *M. pneumoniae* proteins longer than 19 (after *in silico* translation of *M. pneumoniae* genome in the six putative frames) and a list of the common contaminants (599 entries). We set a precursor ion mass tolerance of 15 ppm at the MS1 level and a fragment ion mass tolerance of 0.5 Da. We allowed up to three miscleavages for trypsin. Oxidation of methionine and protein acetylation at the N-terminus were defined as variable modifications, whereas

carbamidomethylation on cysteines was set as a fixed modification. False discovery rates (FDR) in peptide identification were evaluated using a decoy database set to a maximum of 5%.

RNA-seq and shotgun proteomics data analysis

RNA-seq: reads were mapped as explained above to obtain the $\log_2$(CPKM) values. Data from the 9 experiments were quantile-normalized. Each experiment (1 biological replicate, 2 technical replicates) was compared to the rest of experiments, being thus used as internal controls. Comparison was performed two-fold, by calculating the fold-changes in gene expression and performing a t-test between the samples and the internal controls. A multiple-test correction was applied to correct the p-values of the t-test. We only considered as biologically significant those changes with absolute fold-changes larger than 0.8 and corrected p-values smaller than 0.05.

Shotgun proteomics: to obtain reliable protein expression values, only unique peptides (those uniquely mapping to a single protein) were considered. The three largest areas from peptides of the same protein (top 3 peptides) were averaged to obtain a single value for each protein. Areas were rescaled so as to each experiment would have the same expression baseline. Comparison was performed two-fold, by calculating the fold-changes of the areas (in log2) and by performing a t-test between the samples (1 biological replicate; 2 technical replicates for each experiment) and the internal controls (the rest of the samples of the experiment). We applied a multiple-test correction to the p-values of the t-test. Again, we only considered as significant those changes with absolute fold changes larger than 0.8 and corrected p-values smaller than 0.05.

Results

To investigate these hypotheses, we annotated sRNAs *de novo* in the genomes of *Buchnera aphidicola*, *Mycoplasma hyopneumoniae* and *Mycoplasma mycoides* subspecies *capri* (Tables S1-S3; Fig. S3) in a similar way as we did with *M. pneumoniae*[247]. We also considered the sRNAs annotated using deep-sequencing data in other 17 bacterial genomes and a chloroplast genome (Table S4). These 21 genomes span an AT content ranging from 28 to 80%, and their genome sizes range from 416 Kb to 9.02 Mb. Investigating the number of canonical Pribnow boxes in these genomes, we found an exponential dependency of the number of boxes on the AT content, qualitatively similar to our theoretical expectations (Fig S2a). Moreover, comparison of the number of these boxes upstream ORFs and sRNAs showed that the proportion of sRNAs with Pribnow boxes is similar or higher to the proportion of ORFs having them (Fig. S2b). This supports the hypothesis that an increase in AT content also results in an increase in spurious Pribnow boxes.

We found that the number of sRNAs normalized by genome size versus the AT content in the studied bacterial species has a clear exponential dependency (Fig. 1a), similar

to that of the number of TANAAT motifs randomly expected given a certain AT% (Fig S2a). The exponential trend observed for the sRNAs is conserved omitting the species whose sRNAs were de novo annotated (R^2=0.814) indicating that it is not an artifact of the method used to identify them (see Fig. S3 and Methods). In contrast to the observed sRNA trend, the number of coding genes normalized by genome size shows no dependency on AT content, and this trend is invariant with respect to genome size (Fig. 1b). We tested whether the AT dependency held true for both asRNAs and trans-encoded sRNAs. Interestingly, asRNAs follow an exponential dependency on the AT content (Fig. S4a); whereas trans-encoded sRNAs behave similarly to coding genes, and are uncorrelated to the AT content of the intergenic regions (even when considering a minimal size larger than that of an average asRNA; Fig. S4b). These results support the transcriptional noise hypothesis, and that random mutations in coding genes could result in spurious antisense TANAAT boxes, in a manner related to the genome AT content, which could drive the expression of asRNAs.

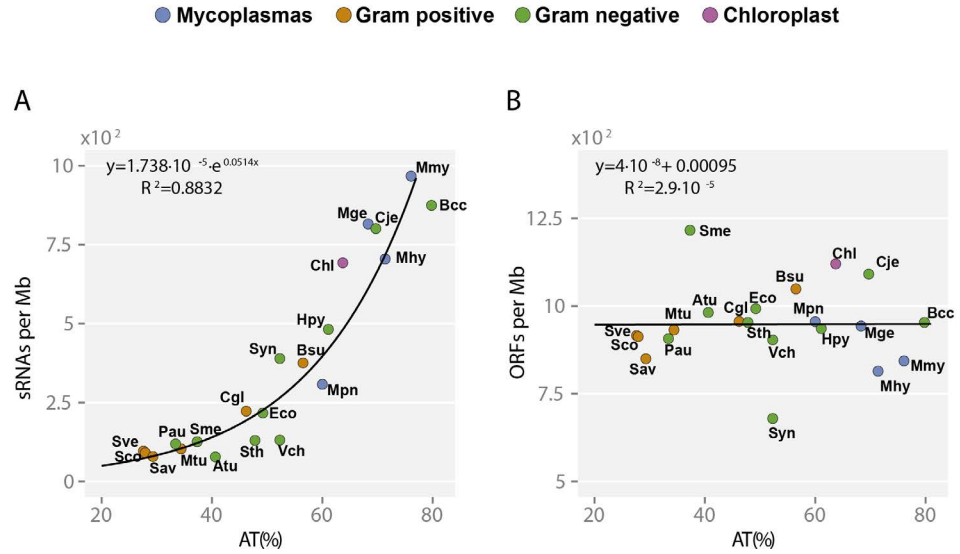


Figure 1. Different genomic features show distinct dependency on the genomic AT content. The number of features was divided by the genome size for normalization and represented versus the genomic AT content. Genomes represented: Atu: *Agrobacterium tumefasciens*; Bcc: *Buchnera aphidicola* (str Cc); Bsu: *Bacillus subtilis*; Cgl: *Corynebacterium glutamicum*; Chl: Chloroplast (*Arabidopsis thaliana*); Cje: *Campylobacter jejuni*; Eco: *Esherichia coli*; Hpy: *Helicobacter pylori*; Mge: *Mycoplasma genitalium*; Mhy: *Mycoplasma hyopneumoniae*; Mmy: *Mycoplasma mycoides*; Mpn: *Mycoplasma pneumoniae*; Mtu: *Mycobacterium tuberculosis*; Pau: *Pseudomonas aeruginosa*; Sav: *Streptomyces avermitilis*; Sco: *Streptomyces coelicolor*; Sme: *Sinorhizobium meliloti*; Sth: *Salmonella typhimurium*; Sve: *Streptomyces venezuelae*; Syn: *Synechocystis* spp, Vch: *Vibrio cholerae*. A) Number of total sRNAs in different bacteria. Total sRNAs have an exponential dependency on the AT content (R^2=0.88) and do not correlate with genome size. B) Genome compaction (i.e. number of ORFs normalized by genome size) versus AT content. Genome compaction in the different bacterial genomes analyzed shows no dependency on the AT content. Instead, the number of ORFs in bacterial genomes correlates with the genome size (R=0.99).

Regarding expression levels, it has been shown that essential ORFs show higher mRNA levels, suggesting that elements with essential roles are more transcribed[100]. Therefore, we compared transcript levels of ORFs and asRNAs in eight of the bacteria in our study. In all cases, average asRNA levels were lower than average mRNA levels (Fig. S5a). This could indicate that at least a majority of the asRNAs could be non-essential. Indeed, a recent study on the essentiality of *M. pneumoniae* genome revealed that only 5% of all sRNAs are essential[100]. We also compared the expression of each asRNA to its overlapping mRNA. asRNA-mRNA expression ratios are represented in Fig. S5b. These ratios are in the majority of cases below 1 (Fig. S5b). For three of the species in our study (*M. pneumoniae*, *M. mycoides* and *B. subtilis*), we compared asRNA levels at exponential and stationary growth phases (Fig S5c). The majority of asRNAs remain unchanged, excluding the effect of the growth phase at where the bacteria were analyzed. We also compared the asRNA levels with the trans-encoded sRNA levels in five species (*B. aphidicola*, *M. genitalium*, *M.*

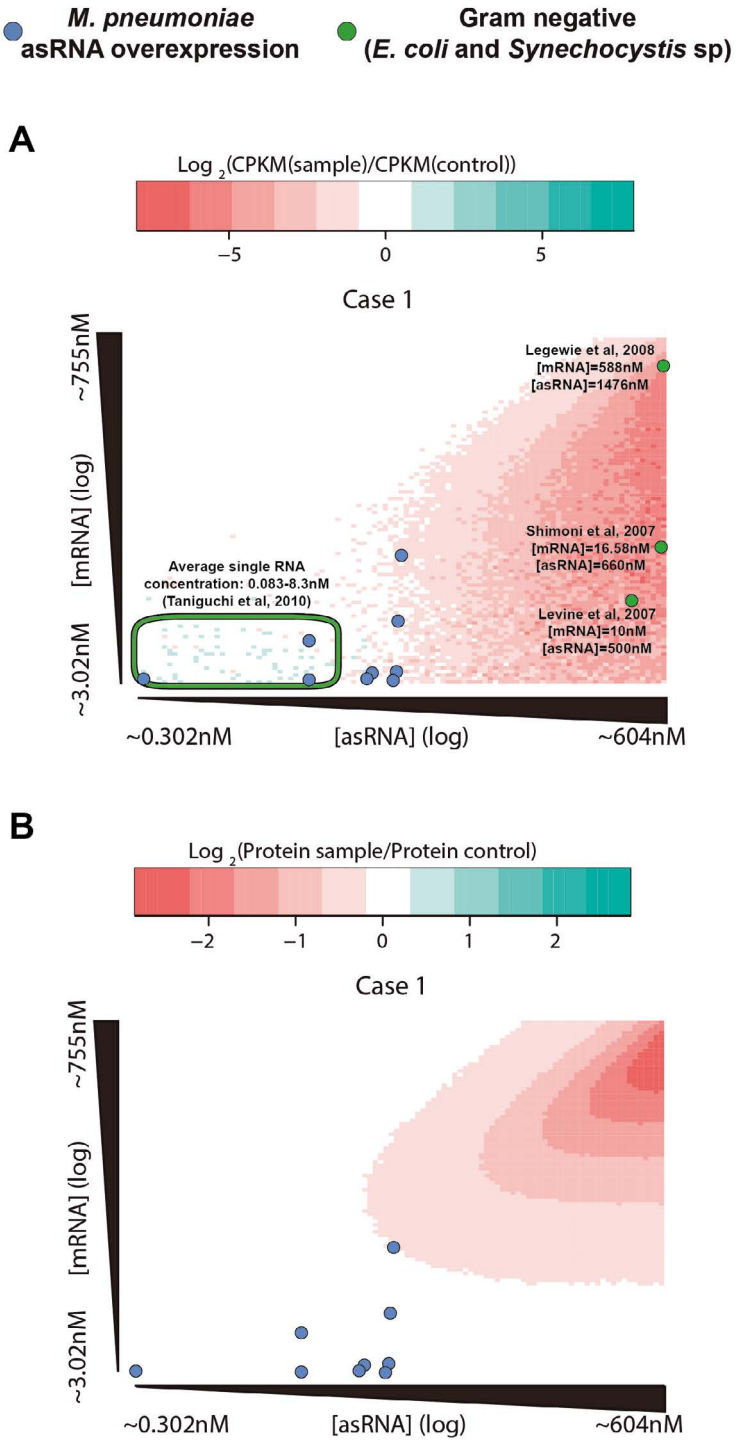
pneumoniae, *M. mycoides* and *M. hyopneumoniae*), and found that in all cases asRNA expression is significantly lower than trans-encoded sRNA levels (Welch's Two Sample t-test, p-value < 0.05).

We estimated the energy consumed by the cells in transcribing these asRNAs in *M. pneumoniae*, considering the number of non-coding RNAs, their length and their transcription rate, compared to those of mRNAs, tRNAs and rRNAs (See Methods). *M. pneumoniae* spends ~5000 ATP units per cell per second in transcribing mRNAs, tRNAs and rRNAs[67]. This amount is proportional to the transcription rate of these molecules, their length, and their number. Taking into account these parameters for sRNAs, we estimate that *M. pneumoniae* spends 2.94% of the energy of RNA transcription in synthesizing sRNAs, equivalent to ~147 ATP units per cell per second. This number represents 0.24% of the total ATP generated per cell per second[67]. Thus, according to our calculations, the energetic impact of spurious transcription is not high even in bacteria with a large number of asRNAs.

asRNAs have been proposed to play a role in transcription regulation complementing the role of transcription factors[248]. Should this be the case, we would expect a negative dependency with the number of transcription factors in the different bacteria analyzed here. The number of transcription factors, as reported in the P2TF database[249], shows a linear trend with genome size as previously described[250] (Fig. S6a). However, this trend does not exist for asRNAs (Fig. S6b). To determine if there is a negative dependency between transcription factors and asRNAs, we considered groups of genomes with approximate similar AT content and different number of transcription factors. We found no negative relationship between the number of transcription factors per genome and the number of asRNAs per genome having similar AT content (>60%) (Fig. S6c). In fact, for bacteria with high AT content, there is a positive correlation, contrary to what we would expect (R=0.94). This can be explained by the fact that for this group, larger genomes present both more transcription factors and asRNAs, Indeed, for bacteria with similar AT content, number of asRNAs correlates with the number of genes, an indicative for genome size (Fig. S6d).

As we indicated in Fig. S1b, asRNAs expressed at low levels could barely encounter its sense mRNA, given the stochastic nature of transcription. Therefore, no effect on mRNA half-life or translation would be expected. To see if this is the case, we constructed a mathematical model of transcription and translation of a gene in the bacterium *M. pneumoniae*. We modeled three possible effects of the asRNA: i) the binding of the asRNA to the mRNA induces degradation of the duplex, ii) the binding of the asRNA to the mRNA induces degradation of the mRNA, and iii) the binding of the asRNA to the mRNA is stable, but prevents translation (Fig. S1a). In all cases, binding of the mRNA to the ribosome prevents degradation of the mRNA. Parameters for this model were determined from experimental data (see Methods). Other possible effects, such as transcriptional interference, were not considered as the low transcription rates in *M.*

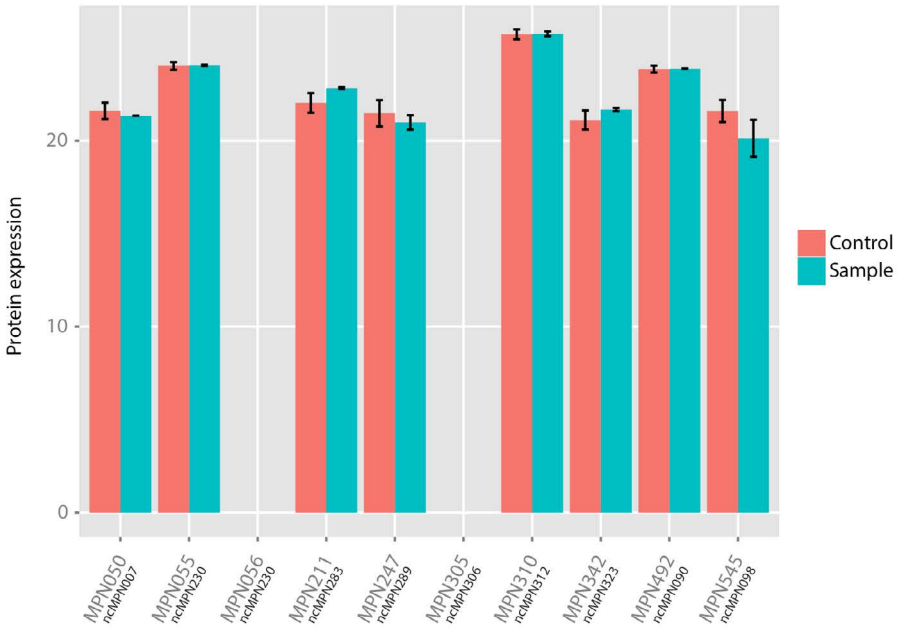
pneumoniae deem very unlikely the collision of transcribing polymerases. We scanned the parameter space of the mRNA and the asRNA transcription rates, from typical wild-type levels to ~100-fold overexpression (Fig. 2 & Fig. S7). Interestingly, we found that for the three cases modeled, the region with low concentrations of both asRNA and mRNA shows no changes with respect to the control simulations. This can be explained by the fact that in this region, RNA copy numbers are below one per cell, and thus the chance of an mRNA and an asRNA to occur simultaneously at the same cell is negligible (Fig. S1b). Remarkably, the majority of RNAs in different bacteria are present at these concentrations that yield no asRNA effect[251], although some exceptions have been described, showing that some asRNAs can have a regulatory role[252–254] (Fig. 2a). This mathematical model can be a valuable resource in order to identify putative functional asRNAs in a given organism according to their expression levels. By determining the concentrations of all asRNAs in *M. pneumoniae*, we can determine a list of potential functional asRNA candidates. In this bacterium, there are no asRNAs naturally expressed enough to potentially trigger an effect in their overlapping mRNAs, according to our simulations. It has to be noted, though, that the values of decay rates used in these simulations represent the average values determined for *M. pneumoniae*. Individual transcripts with decay rates that differ significantly from the average should be analyzed on a case-by-case basis. With the adequate parameters, the model could be extended to other bacteria, given that the action mechanism of asRNAs is known beforehand.


M. pneumoniae
asRNA overexpression
Gram negative
(E. coli and Synechocystis sp)
A
Log 2(CPKM(sample)/CPKM(control))
−5
0
5
Case 1
~755nM
[mRNA] (log)
~3.02nM
Legewie et al, 2008
[mRNA]=588nM
[asRNA]=1476nM
Shimoni et al, 2007
[mRNA]=16.58nM
[asRNA]=660nM
Levine et al, 2007
[mRNA]=10nM
[asRNA]=500nM
Average single RNA
concentration: 0.083-8.3nM
(Taniguchi et al, 2010)
~0.302nM
[asRNA] (log)
~604nM
B
Log 2(Protein sample/Protein control)
−2
−1
0
1
2
Case 1
~755nM
[mRNA] (log)
~3.02nM
~0.302nM
[asRNA] (log)
~604nM

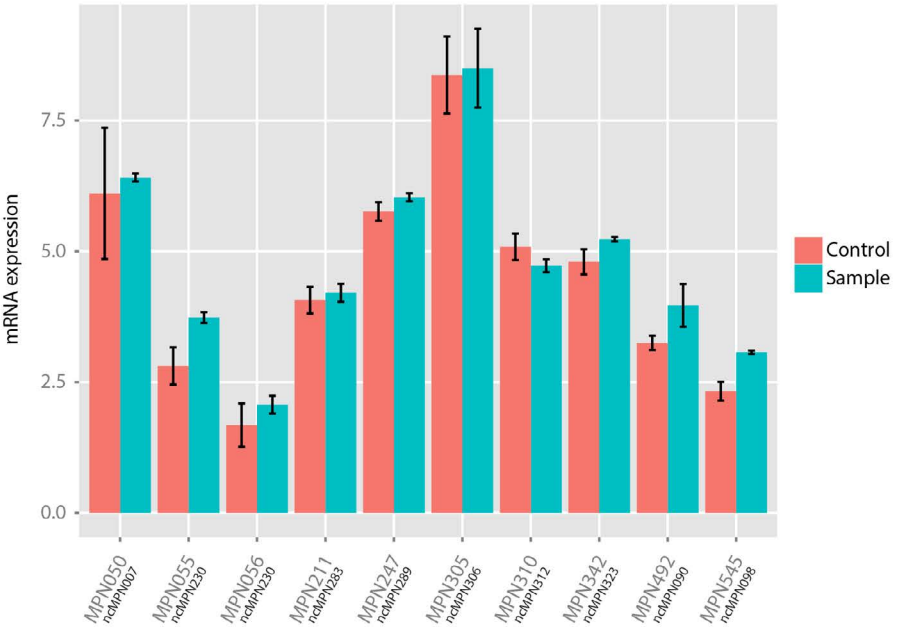
◄ **Figure 2**. Simulation of the effect of the asRNAs, assuming that the pairing asRNA-mRNA causes duplex degradation. Parameters for the simulations are detailed in the Supplementary Information. Each point of the heatmaps represents the average change in the protein concentration for 100 simulations of 1000 minutes each, for specific parameters of asRNA and mRNA transcription rates. The remaining parameters remain constant for all the simulations. The axes represent the mRNA and asRNA concentration in the control experiments for the corresponding transcription rates scanned. A) Changes in the mRNA concentration after 1000 minutes of simulation. Blue circles represent experimental data from the overexpression of asRNAs in *M. pneumoniae*, whilst green circles represent data from studies in Gram – bacteria[252–254]. The green ellipse delimits the region of the concentrations of the majority of transcripts in *E. coli*[251]. B) Changes in the protein concentration after 1000 minutes of simulation. Blue circles represent experimental data from the overexpression of asRNAs in *M. pneumoniae.*

To verify these results, we overexpressed 9 asRNAs in the bacterium *M. pneumoniae*, (up to 6-fold; Fig. 2 and Table S5). These asRNAs were selected so as to overlap different regions of their corresponding mRNA partners (5'-end, 3'-end or center), in order to test different possible action mechanisms. Additionally, asRNAs with different expression levels were chosen. Shotgun proteomics of the clones revealed no significant changes in the protein levels of the overlapping genes (Fig. 3a and Table S6). Also, RNA-seq revealed no significant changes in the mRNA levels (Fig. 3b and Table S7). Thus, our simulations and our experimental data do not support the hypothesis that asRNAs have a general regulatory role in bacteria replacing the function of transcription factors. Only in those exceptions in which both asRNA and mRNA are both expressed over a certain threshold, a regulatory behavior can be expected.

A


Protein expression
0
10
20
Control
Sample
MPN050 ncMPN007
MPN055 ncMPN230
MPN056 ncMPN230
MPN211 ncMPN283
MPN247 ncMPN289
MPN305 ncMPN306
MPN310 ncMPN312
MPN342 ncMPN323
MPN492 ncMPN090
MPN545 ncMPN098

B


mRNA expression
0.0
2.5
5.0
7.5
Control
Sample
MPN050 ncMPN007
MPN055 ncMPN230
MPN056 ncMPN230
MPN211 ncMPN283
MPN247 ncMPN289
MPN305 ncMPN306
MPN310 ncMPN312
MPN342 ncMPN323
MPN492 ncMPN090
MPN545 ncMPN098

◄ **Figure 3.** Effect of the overexpression of asRNAs in their overlapping genes, measured by RNA-seq and shotgun proteomics. A) Protein levels of the genes overlapping each asRNA in control conditions and in the strains transformed with the antisense constructs. Error bars represent the standard deviation of the samples. Two of the proteins, MPN056 and MPN305, were not detected in any of the strains of *M. pneumoniae*. B) mRNA levels of the genes overlapping each asRNA, in control (wild-type) conditions and in the strains overexpressing the antisense transcripts. Error bars represent the standard deviation of the samples.

Our findings support the idea that many of the asRNAs are consequence of transcriptional noise, rather than of tightly regulated events. The distribution of asRNAs in bacteria with distinct AT content and the lack of capability of replacing transcription factors support this idea. Probably, the bias towards AT mutations in bacteria[233] generates spurious promoter sequences that are able to trigger transcription. However, spurious expression of asRNAs is not incompatible with some of them being functional, as described elsewhere[214,215,219–221,224]. Indeed, asRNAs claimed to be functional are expressed at much higher rates than the average[251–254]. Despite the observed general trend, we should not ignore that in some bacteria, there are proteins (such as RNA chaperone Hfq[255]) that help to stabilize asRNAs or the duplexes they form with mRNAs. In such cases, even low expressed asRNAs may exert a regulatory function. Nevertheless, this protein is not conserved throughout the bacteria in our study; and in some species, although conserved, it is not essential. Therefore, we cannot expect such a mechanism to be general but an adaptation for specific cases. This suggests that asRNAs may accumulate in bacterial genomes due to noisy transcription and a lack of negative selection, probably due to the low energy needed for their transcription and the absence of deleterious effects. Some of these asRNAs may afterwards gain a function. Additionally, pervasive non-coding transcription may as well have unspecific functional roles, such as buffering the RNA polymerase levels inside the bacterial cell. Our results are likely to be valid throughout the bacterial kingdom, and according to a recent study[256], they may also apply to eukaryotes.

Acknowledgements

We thank the Genomics, Proteomics and Protein Technologies Core Facilities at CRG. This work was supported by the European Union Seventh Framework Programme (FP7/2007–2013), through the European Research Council [232913]; Fundación Botín, the Spanish Ministry of Economy and Competitiveness [BIO2007-61762]; National Plan of R + D + i; ISCIII – Subdirección General de Evaluación y Fomento de la Investigación [PI10/01702]; European Regional Development Fund (ERDF) (to the ICREA Research Professor L.S.]; Spanish Ministry of Economy and Competitiveness, 'Centro de Excelencia Severo Ochoa 2013–2017′ [SEV-2012-0208]. AL received grant BFU2012-39816-C02-01 from the Spanish Ministry of Economy and Competitivity co-financed by FEDER funds.

Chapter 7

General discussion

Mycoplasma hyopneumoniae (*M. hyopneumoniae*) causes enzootic pneumonia in pigs and is a major contributor to development of the respiratory disease complex[2]. Bacterin vaccines offer protection against clinical signs caused by *M. hyopneumoniae* infections, but the specific components needed to elicit protective immunity are unknown. Also, the growth requirements of *M. hyopneumoniae* are largely unknown, necessitating the use of complex cultivation media containing components from porcine and bovine origin. To improve the robustness of the production process for *M. hyopneumoniae* bacterin production a thorough understanding of the growth and survival strategies of *M. hyopneumoniae* is needed which requires in-depth analysis of the metabolic capabilities of this bacterial pathogen. In this thesis we created a genome-scale metabolic model (GEM) of *M. hyopneumoniae* to determine *in vitro* flux distributions in aerobic fermentor systems, analyzed differentially expressed (metabolic) genes during infection and studied metabolic functions in relation to host specificity (figure 1). By model-driven experimentation we were able to increase biomass yield 2.3 times by addition of pyruvate to the cultivation medium. To better understand the need for porcine components in the cultivation medium we performed a functional comparison of 80 mycoplasma species based on the protein domain composition and identified functional differences between mycoplasma species based on phylogeny and host. Growth strategies were further elucidated by sequencing the *in vivo* transcriptome and identification of differentially expressed genes and ncRNAs during infection when compared to culture grown cells. This analysis showed *in vivo* up-regulation of genes needed for glycerol-3-phoshate and spermidine transport, nucleotide metabolism and an F1-like ATPase that is likely needed to maintain pH homeostasis (figure 1). Overall, this study provided novel insight in growth and survival strategies of *M. hyopneumoniae* under *in vitro* and *in vivo* conditions and is the first study which captured the *in vivo* transcriptional landscape using RNA sequencing. Knowledge on metabolism was captured in a genome-scale metabolic model which provides the basis for model-driven experimentation. However, the systems analysis performed in this thesis did not yet provide a more defined growth medium nor verified the role of potential virulence factors during infection. Next steps to increase our understanding of *M. hyopneumoniae* from a systems perspective should follow from model pathway validation using 13C-labelled components, further analysis of transcriptional flexibility in varying environments and generation of transposon mutants to study the role of potential virulence factors and verify essential genes needed for growth in cultures or in the host.

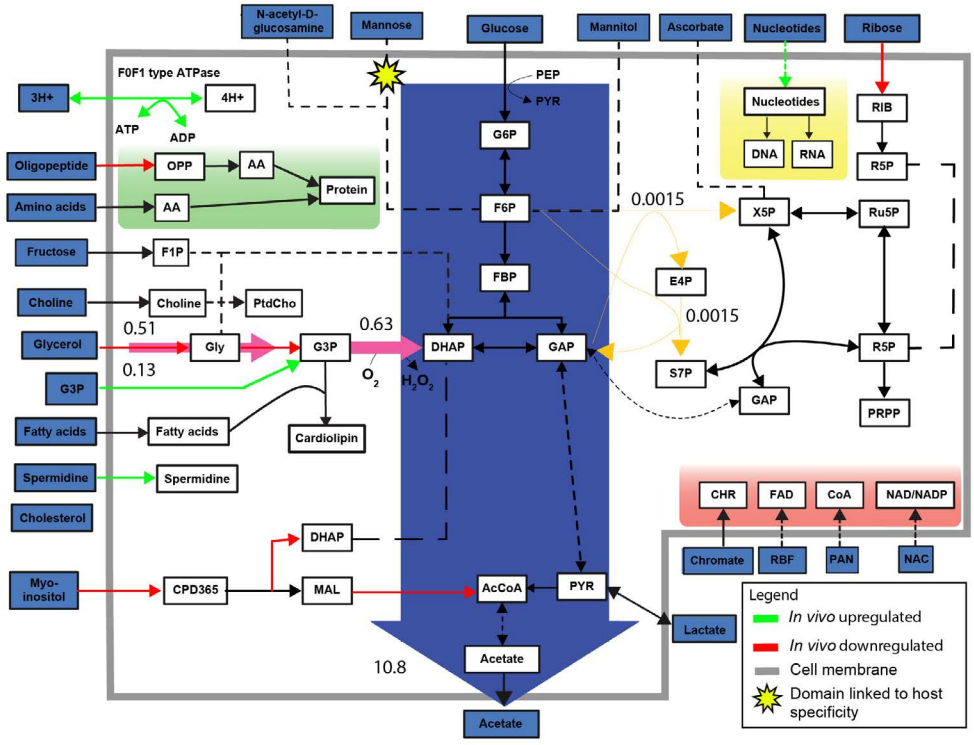


Figure 1. ***M. hyopneumoniae* metabolic adaptation to *in vivo* and *in vitro* environments.** Metabolic map of *M. hyopneumoniae* showing flux distribution in cultures derived from the metabolic model (chapter 3), differentially expressed genes during infection derived from chapter 5 and an enzyme with a domain linked to host specificity derived from chapter 3. Flux through the main metabolic pathways are shown, blue (glycolysis), pink (glycerol) and orange (pentose phosphate pathway), numbers show fluxes in mmol·gDW^{-1}·h^{-1}. Differentially expressed metabolic enzymes during infection are shown in green and red (not scaled according to flux). Mannose-6-phosphate isomerase contains an RmlC-like jelly roll fold (IPR014710) which was found linked to host specificity in chapter 4. Abbreviations are explained in table S11 of chapter 3.

As discussed in the first chapter of this thesis, there is a great scientific interest in mycoplasma species and recent publications have provided novel insights into growth and survival strategies used by *M. hyopneumoniae* that were not yet fully incorporated in our GEM. In this final chapter, I will discuss how the findings in this thesis and findings from recent literature can be applied to improve the *M. hyopneumoniae* GEM. Four novel system strategies (figure 2) used by *M. hyopneumoniae* will be discussed: i) production of surface proteins and extracellular proteins, ii) production of oxygen radicals, iii) alternative metabolic pathways and iv) alternative transcriptional landscapes and transcriptional regulation. After discussing these topics, I will discuss if our findings in *M. hyopneumoniae* apply to other mycoplasma species and indicate overlaps and species differences. Finally, I will provide future directions for process development of *M. hyopneumoniae* bacterin vaccines.

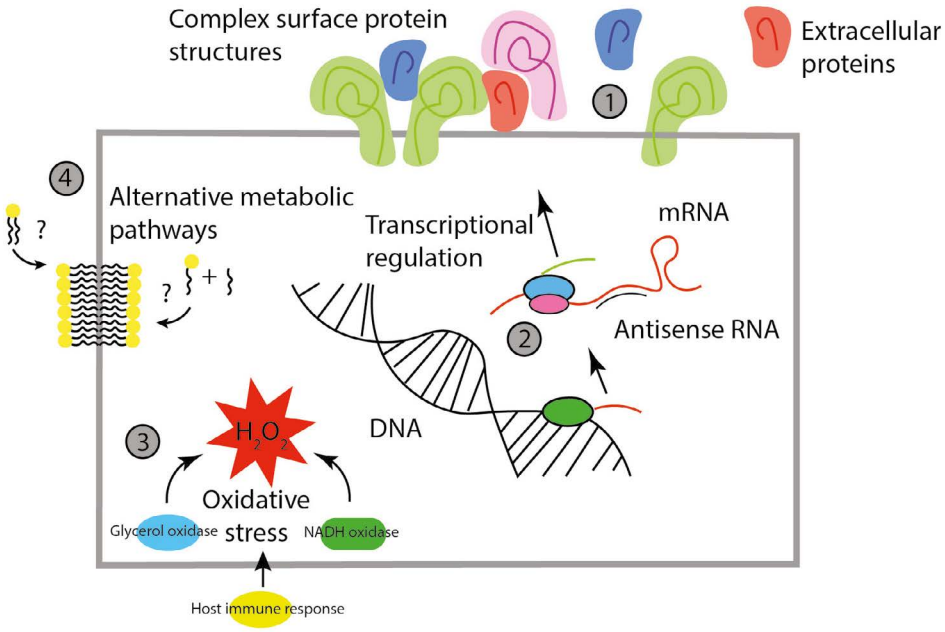


Figure 2. **Alternative system properties of *M. hyopneumoniae* to be incorporated in the GEM.** Shown are: 1) complex surface protein structures and secreted proteins, 2) transcriptional regulation, 3) oxidative stress and 4) alternative metabolic pathways.

Production of surface proteins and extracellular proteins

The generalized biomass equation that was used as a basis for our GEM (chapter 3) was already determined in 1963 by Razin *et al.* for *M. bovigenitalium*[127] which is phylogenetically closely related to *M. bovis* and *M. fermentans*[257]. Recent literature provided novel insight on post-translational protein modification and protein translocation in *M. hyopneumoniae* which increases our understanding of the protein composition of biomass. The *M. hyopneumoniae* genome contains genes needed to perform basic protein translocation functions such as *secA*, *secY*, *secD*, *prsA*, *dnaK* and *lepA*[15]. Notably absent are *secB*, *secD*, *secE*, *secG*, *secF*, *SPaseI* and the *groEL-groES* operon. Although the protein translocation machinery is limited, *M. hyopneumoniae* is able to express various lipoproteins and general membrane proteins at the cell surface. A review of the protein translocation mechanisms in mycoplasma species was recently provided[258], here we will only discuss impact on the metabolic model of protein modification and translocation.

In **chapter 5** we found that genes related to cilium adhesion were highly expressed in *M. hyopneumoniae*. Recent studies found that besides these specific cilium adhesins, proteins normally expected to be present in the cytosol were able to bind host-derived components (e.g. fibronectin and heparin) and were found present at the cell surface,

indicating that these proteins have a secondary function facilitating adhesion[185,259]. In addition, many proteins present at the cell surface were proteolytically cleaved in a specific manner[29,32,182,183]. This leads to the hypothesis that some (highly expressed) cytosolic proteins in *M. hyopneumoniae* have an additional role in cilium adhesion. Some of the protein fragments created at the cell surface have no direct chemical attachment to the cell membrane and it is expected that these proteins remain attached to the cell membrane via hydrophobic interactions with other surface proteins, interactions with components derived from the host or via yet unknown active mechanisms[182]. It is well established that mycoplasma species create diversity at the cell membrane by expressing variable lipoproteins but the creation of these complex protein structures (figure 2) could be a unique trait of *M. hyopneumoniae* and could influence the protein composition of biomass used in the GEM because the build-up of an extracellular bacterial protein matrix was not taken into account. In the model this can be incorporated by increasing the percentage of protein in the biomass equation. It should also be verified if these extracellular protein structures are essential for growth and whether host-derived components play an essential role in the formation of the surface structures. If this is the case one or multiple host-derived components (e.g. fibronectin or heparin) or components with an equivalent function should be added to the growth medium used for production of bacterin vaccines.

Although it is assumed that it is in the best interest of the bacterium to keep the cilium adhesion protein fragments attached to the cell surface, there is a risk that protein fragments are lost in the extracellular environment (figure 2). Only recently the protein secretome was determined for *M. hyopneumoniae*[260] and was found to consist of more than 60 proteins, approximately 9% of the predicted protein coding sequences in the *M. hyopneumoniae* genome. Production of extracellular proteins was not taken into account in our genome-scale metabolic model (**chapter 3**) but could be introduced by defining a new reaction for the production of extracellular protein. This obviously would increase the total energy expenses for protein synthesis and may provide an additional explanation for the high percentage of energy needed for non-growth associated maintenance (**chapter 3**). Further quantitative understanding is needed of post-translational modifications, translocation and secretion of *M. hyopneumoniae* proteins and this knowledge should be implemented in the GEM.

Production of oxygen radicals

The capability to produce oxygen radicals, such as hydrogen peroxide[36] or hydrogen sulfide[140], is described as an important virulence factor for mycoplasma species because these compounds cause damage to host tissue and could release nutrients for growth. Production of these compounds may also cause endogenous oxidative stress in mycoplasma species and protective mechanisms are needed to prevent damage to bacterial proteins, DNA and cell membrane lipids. Furthermore, during infection, *M. hyopneumoniae* will

experience oxidative stress as a result of radical production by alveolar macrophages and neutrophils (figure 2). Our GEM (**chapter 3**) contains two reactions that could generate hydrogen peroxide and cause endogenous oxidative stress: glycerol oxidase and NADH oxidase. Studies in other mycoplasma species have shown that the amount of hydrogen peroxide produced by these enzymes varies per species and depends on the substrate oxidized[261]. In general, the oxidation of glucose results in a relatively low yield of hydrogen peroxide production per substrate while the oxidation of glycerol or glycerol-3-phosphate results in a high yield of hydrogen peroxide per substrate. Only two enzymes were described in *M. hyopneumoniae* that could protect against oxidative stress: a peroxidase (*tpx*)[262,263] and thioredoxin reductase[169]. Alternative protective enzymes, catalase, glutathione peroxidase or superoxide dismutase were not found in the *M. hyopneumoniae* genome[16,264]. Besides protective enzymes, components obtained from the host or growth medium could offer protection against oxidative stress. The cell membrane of Mycoplasma species contains a high concentration of cholesterol[14] which could prevent peroxidation of polyunsaturated fatty acids in cell membrane lipids[265]. Furthermore, ascorbic acid which can be metabolized by *M. hyopneumoniae* (**chapter 3**) could play a role in prevention of oxidative stress, although exact mechanisms are unknown. In our analysis of the *in vivo* transcriptome (**chapter 5**) we found up-regulation of DNA polymerase I and excinuclease subunit A and we hypothesized that this was possibly related to oxidative stress during infection. Also, we identified up-regulation of glycerol-3-phosphate permease and downregulation of glycerol transport which could influence flux through the glycerol oxidase and increase or lower the hydrogen peroxide production rate.

The response of *M. hyopneumoniae* to oxidative stress was studied using microarrays[266]. Up-regulated genes were mainly hypothetical indicating that we do not have sufficient knowledge to understand the response of *M. hyopneumoniae* to oxidative stress. Down-regulation was observed of neutrophil-activating protein (*napA*). An orthologous gene in *H. pylori* has a double role during infection since it induces neutrophils to produce reactive oxygen species and promotes neutrophil adhesion to epithelial cells but also binds to DNA and offers protection against oxidative damage[186,267]. The role of this protein in *M. hyopneumoniae* during infection remains to be elucidated. Overall, the response to oxidative stress was moderate with low fold changes in gene expression levels and a limited number of differentially expressed genes. Further investigation of the influence of oxidative stress on the growth of *M. hyopneumoniae* in culture systems and in the host is needed. In the GEM, the influence of endogenous hydrogen peroxide production could be represented by including an ATP penalty depending on the amount of hydrogen peroxide produced.

Alternative metabolic pathways

The focus of our modeling approach was on energy balances and assignment of gene-protein-relations was done based on the presence of functional protein domains. Novel metabolic pathways or reactions could be found in the near future by further exploration of the functions of hypothetical proteins in *M. hyopneumoniae*. Elucidation of the functionality of novel enzymes and incorporation of these functions in the GEM should further increase our understanding of the metabolic capability of *M. hyopneumoniae*. Here, the lipid metabolic pathways and myo-inositol pathway in *M. hyopneumoniae* will be further discussed.

Lipid synthesis capabilities in our model are minimal and not sufficient to produce the lipids found in *M. hyopneumoniae* biomass cultured on complex medium: sphingomyelin, phosphatidylcholine and diphosphatidylcholine (cardiolipin)[268]. Only cardiolipin can be produced according to the model. For production of phosphatidylcholine and sphingomyelin essential reactions are missing. During his master thesis project, Luuk van Schijndel investigated whether the enzyme needed to form phosphatidylcholine from fatty acids and sn-glycero-3-phosphocholine, phospholipase B (*plb*) could be detected in *M. hyopneumoniae* cell membranes, as this function was present in the GEM of *M. pneumoniae*[67]. He used a hydroxamate assay but no activity was found and this specific reaction could not be added to the genome-scale metabolic model of *M. hyopneumoniae* (data not shown). Alternative synthesis reactions should be considered or a more sensitive assay is needed. In a follow-up study, we measured the lipid composition of *M. hyopneumoniae* biomass and identified sphingomyelin, phosphatidylcholine and phosphatidylethanolamine as major phospholipids using APCI-MS and thin-layer chromatography. The measured lipidome of biomass was complex with more 500 different lipids and resembled the lipidome of porcine serum (data not shown). Based on these measurements the minimal lipid composition of *M. hyopneumoniae* could not be determined and a basic lipid composition was assumed for the model consisting of cardiolipin, phosphatidylcholine and phosphatidate (**chapter 3**). Sphingomyelin and cholesterol as lipid components were not taken into account in the biomass equation used in the GEM, although these components were measured, they cannot be produced by the bacterium and will not influence the intracellular energy balance. Incorporation of cholesterol should be done when the GEM is used to design growth medium because this component is needed for membrane integrity and is present in all defined mycoplasma media[75,112,269]. Whether sphingomyelin is needed in biomass or is just present because it is present in serum remains to be determined. Sphingomyelin and cholesterol have been described to form lipid rafts in cell membranes with specific functionalities regarding distribution of cell membrane components and possibly host invasion[270,271]. Whether lipid rafts are important for growth or virulence of *M. hyopneumoniae* remains to be determined.

Synthesis of glycolipids was not incorporated in our GEM although monogalactosyldiacylglycerol (MGDG) and digalactosyldiacylglycerol (DGDG) were found in lipid extracts from *M. hyopneumoniae* membranes isolated from strains M and Z[272]. Further studies are needed to determine if these glycolipids are also produced by the strains used in this thesis. It is likely that glycolipid synthesis is essential for growth and needed for membrane structural properties[273], as even the minimal synthetic organism JCVI-Syn3.0 contains a glycolipid synthesis pathway[86], although, as we show in **chapter 4**, alternative functionalities could exist that by-pass these essential functions. If glycolipids play an essential role in biomass, the reactions used to form glycolipids should be added to the GEM.

Another pathway which requires further attention is the myo-inositol pathway. Genes related to myo-inositol metabolism were highly expressed (**chapter 5**) *in vitro* and *in vivo* and most likely provide the cell with an additional carbon source, although myo-inositol consumption was not measured in *M. hyopneumoniae* cultures (**chapter 3**) and a specific transporter was not found in the genome (**chapter 3**). It should be further investigated if myo-inositol can be consumed from phosphatidylinositol present in pulmonary surfactant[274]. Furthermore, an alternative function for the enzymes in the myo-inositol pathway could exist, not directly related to myo-inositol catabolism. For instance, methylmalonate-semialdehyde dehydrogenase, which is part of the myo-inositol pathway was found present on the cell surface[185] and could have a role in cilium adhesion[259]. Alternatively, this protein has been shown to interact with RNA and binds NAD(P)H[275] which could mean it functions as a scavenger for these compounds.

Mechanisms for transcriptional regulation in *M. hyopneumoniae* need to be further elucidated

In this thesis project, we applied RNA sequencing to study the transcriptome of *M. hyopneumoniae* during exponential growth in cultures and during infection (**chapters 3, 5 and 6**). Overall, 3 different culture conditions were tested with two different strains and *in vivo* growth was studied in tissue and in lung flush samples. In *M. pneumoniae* insight into transcriptional regulation was obtained by measuring transcript levels in 115 different growth conditions and it was shown that co-directional adjacent operons could be transcribed into one mRNA depending on the growth conditions[56]. Furthermore, it was shown that transcription could be repressed by DNA regions preferentially bound by RNA polymerases, intrinsic terminators and large intergenic regions. Whether such mechanisms also play a role in *M. hyopneumoniae* will require additional datasets collected using varying growth conditions. One of the conditions that would be interesting to study is the transcriptional response to starvation since we found *in vivo* down-regulation of genes related to cell division (**chapter 5**). Furthermore, in cultures grown using complex medium

it was not clear whether shortage of a component in the medium or production of a high concentration of byproducts (e.g. acetate) caused cultures to reach stationary phase (**chapter 3**). Sequencing cultures in stationary phase or starvation could further explain why cultures stop growing and whether the transcriptional changes observed *in vivo* are related to a change in the growth rate or are found as a result of actions of the immune system or changed environmental properties such as temperature or nutrient availability. When the transcriptome of *M. hyopneumoniae* is determined under multiple growth conditions and further insight is gained in regulation of gene expression, the different transcriptomes can be analysed using the GEM to better understand the effect of environmental perturbations on cellular metabolism[276–278].

In **chapter 6** we annotated ncRNAs in *M. hyopneumoniae* using a transcriptome dataset collected from a culture grown in fermentor systems (**chapter 3**). We subsequently used the annotated locations for ncRNAs to analyse whether there is differential expression of ncRNAs in *M. hyopneumoniae* during infection (**chapter 5**). Many of these ncRNAs are expected to be the result of spurious translation but some might be essential as was found in *M. pneumoniae*[100]. When additional transcriptome data is collected of *M. hyopneumoniae* this should be done in a strand-specific manner to be able to analyse the amount and distribution of ncRNAs and further analyse which ncRNAs have a regulatory role. This will enable an improved annotation of the ncRNAs in *M. hyopneumoniae* based on more datasets than used in **chapter 6**. Further interpretation of the effects of gene regulation on metabolic fluxes using the genome-scale metabolic model or eventually incorporation of this knowledge into a dynamic model of *M. hyopneumoniae* will be an exciting next step to better understand growth and physiology of *M. hyopneumoniae*[279,280].

Can we relate findings in this thesis to other Mycoplasma species?

We started this thesis with the observation that the scientific interest in Mycoplasma species was high, but mainly focused on the human pathogens *M. pneumoniae* and *M. genitalium*. During this thesis we gained novel insight on mycoplasma functional capabilities and *M. hyopneumoniae* genomics, transcriptomics and metabolomics (figure 1). Some of the findings in this thesis will be unique for *M. hyopneumoniae*, for instance, the findings related to myo-inositol metabolism[116] which is uniquely present in *M. hyopneumoniae* (**chapter 3,5**). To assess whether findings in this thesis will apply to other mycoplasma species we can use the functional comparison provided in **chapter 4** of this thesis.

We showed that the core domainome covered only 19.3% of the pan-domainome of mycoplasma species, indicating that the functional differences between mycoplasma species are large and therefore findings in one species cannot be directly related to another mycoplasma species. We also showed that the evolutionary persistence of hypothetical proteins is low so the differentially expressed hypothetical proteins found in **chapter 5** during infection will probably be uniquely present in *M. hyopneumoniae* and possibly play

a unique role during infection in *M. hyopneumoniae*. Protein domains with a metabolic function were found more conserved than hypothetical proteins in mycoplasma species (33.8% present in the core) and therefore there is a higher chance that findings related to metabolic functions also apply to other mycoplasma species. For instance, the finding in **chapter 5** that nucleases could play an important role during infection is expected to apply for multiple mycoplasma species[48,281–285]. As a result of the overlap in metabolic capability, we expect that our finding of pyruvate as an alternative energy source for *M. hyopneumoniae* will also be relevant for other mycoplasma species as already described for *M. agalactiae* and *M. mycoides*[134]. Some of the findings in this thesis could represent general mechanisms in the Hominis phylogenetic group. For instance, the F1-like ATPase was identified in other mycoplasma species in the Hominis group[206] and could play a more general role during infection although this should be further investigated. We also expect that all mycoplasma species will express a high amount of ncRNAs as the amount of ncRNA is related to the AT content of the genome sequences which is high in all mycoplasma species (**chapter 6**). In general, based on the findings in **chapter 4**, it is expected that findings related to hypothetical proteins will probably be only relevant for *M. hyopneumoniae*, findings related to core (metabolic) functions will be of interest for all mycoplasma species and findings related to core functions in the Hominis phylogenetic group will be of interest for this phylogenetic cluster.

Further improvement of the vaccine production process

The systems analysis presented in this thesis increased our understanding of *M. hyopneumoniae* on multiple levels. We were able to increase growth rate and yield of the cultivation step (**chapter 3**) and identified genes possibly relevant for virulence (**chapter 5**). Next steps needed for improvement of the vaccine production process are: i) validation of pathways in the metabolic model using ^{13}C labeled components, ii) determination of novel critical process parameters using transcriptome datasets, iii) validation of model predicted essential genes by making a mutant library, iv) validation of model-based gene knock-in mutants and v) development of defined cultivation medium.

To better understand *M. hyopneumoniae* growth in the fermentor, the metabolic pathways in the metabolic model should be validated using ^{13}C-labelled components. The most interesting pathways to validate are the pathways used to synthesize lipids and NAD(P)H because these components are needed for growth but genes are missing for essential reactions in the pathways[116]. When the role of these pathways is better established, transcriptome profiles collected under different growth conditions can be correlated to understand the relation with growth rate and the formation of specific antigens (**chapter 2**). Together with data related to, for instance, the quality of the raw materials or on-line data measured during the fermentation, novel critical process parameters could be found using multivariate data analysis (**chapter 2**). Further model validation should be done by

determining essential genes in *M. hyopneumoniae* using random transposon mutagenesis. Currently, only a very limited dataset of non-essential genes in *M. hyopneumoniae* is available[133] and there is a need for a better method to transform *M. hyopneumoniae* using transposon mutagenesis. When a mutant library for *M. hyopneumoniae* is available it could be used to challenge pigs and study essential genes, ncRNAs or essential domains for survival in the host. Gene knock-ins predicted by the model could be verified when more advanced genome-editing techniques have been developed in *M. hyopneumoniae*. Recently, expression of GFP in *M. hyopneumoniae* was accomplished with a plasmid incorporating the *oriC* sequence and GFP expression by the P97 promoter[286]. Such a system could be applied to express a transporter for myo-inositol in *M. hyopneumoniae* and verify one of the gene knock-ins described in **chapter 4**. When genetic tools are available to make targeted gene deletions in *M. hyopneumoniae* a genome-streamlining attempt could be pursued to engineer a more robust production strain for *M. hyopneumoniae* vaccines as described in **chapter 2**.

The systems analysis performed did not allow design of chemically defined media for *M. hyopneumoniae* and cultivation for vaccine production still requires complex media with animal components. The failure to design chemically defined media is partly due to the large number of hypothetical proteins in the *M. hyopneumoniae* genome with yet unknown (metabolic) functions. Also, there was a lack of focus on strictly physiological phenomena that influence growth. As described in this final chapter, cholesterol is needed for growth but with a systems strategy strictly focused on functional domains in the genome, essentiality of this component will not be found because there is no gene with a function related to cholesterol metabolism. We currently lack the knowledge to predict which other components will be needed directly from the medium to form, for instance, a functional mycoplasma cell membrane. As we discussed complex lipids could play a role but for defining a growth medium for *M. hyopneumoniae* we will need to know which of the more than 500 lipids currently present in the membrane are needed, which requires modeling of membrane structures to better understand which components are critical[287,288].

Outlook

Although not all challenges related to *M. hyopneumoniae* bacterin production were solved by the studies described in this thesis, I'm optimistic that with the basis provided in this thesis we will be able to take the next steps to further improve the vaccine production process for *M. hyopneumoniae* bacterins. Some of the future work described, such as the generation of transposon mutant libraries will be pursued in the MycoSynVac project[289] and will allow us to assess which genes are essential for growth in culture medium and provide novel insight in genes needed for virulence. Next steps to validate the GEM are incorporated in an internal project at MSD-AH. Furthermore, the rapid developments in the field of transcriptomics, proteomics and metabolomics will allow us to generate the datasets

needed to better understand regulation of gene expression and investigate if there is a relation between antigenic mass formation and metabolic flux distribution. Also, the further development of bioinformatics tools such as the SAPP pipeline used in this study will allow researchers to more easily process the data generated from high-throughput omics experiments and relate the data to biologically relevant questions. Finally, as discussed in this chapter there should be specific attention for physiological processes not depending on the gene content of the bacterium. Combining systems level understanding with physiological critical process parameters will result in a robust vaccine production process for *M. hyopneumoniae* bacterins.

Bibliography

1. Chae, C. Porcine respiratory disease complex : Interaction of vaccination and porcine circovirus type 2 , porcine reproductive and respiratory syndrome virus , and Mycoplasma hyopneumoniae. *Vet. J.* **212,** 1–6 (2016).
2. Brockmeier, S. L., Halbur, P. G. & Thacker, E. L. in *Polymicrobial Diseases* (eds. Brogden, K. & Guthmiller, J.) 231–258 (American Society for Microbiology, 2002).
3. USDA. *Swine 2012 Part II: Reference of Swine Health and Health Management in the United States, 2012*. (2012).
4. Dee, S., Otake, S., Oliveira, S. & Deen, J. Evidence of long distance airborne transport of porcine reproductive and respiratory syndrome virus and Mycoplasma hyopneumoniae. *Vet. Res.* **40,** 1–13 (2009).
5. Maes, D. *et al.* Control of Mycoplasma hyopneumoniae infections in pigs. *Vet. Microbiol.* **126,** 297–309 (2008).
6. Gonyou, H. W., Lemay, S. P. & Zhang, Y. in *Diseases of Swine* (eds. Straw, B. E., Zimmerman, J. J., D'Allaire, S. & Taylor, D. .) 1027–1036 (Blackwell Publishing Ltd., Oxford, UK, 2006).
7. Maes, D. *et al.* Effect of vaccination against Mycoplasma hyopneumoniae in pig herds with an all-in/all-out production system. *Vaccine* **17,** 1024–1034 (1999).
8. Maes, D. *et al.* The effect of vaccination against Mycoplasma hypopneumoniae in pig herds with a continuous production system. *Zentralbl. Veterinarmed. B* **45,** 495–505 (1998).
9. Thacker, E. L., Thacker, B. J., Kuhn, M., Hawkins, P. A. & Waters, W. R. Evaluation of local and systemic immune responses induced by intramuscular injection of a Mycoplasma hyopneumoniae bacterin to pigs. *Am. J. Vet. Res* **6,** 1384–1389 (2000).
10. Thacker, E. L., Thacker, B. J., Boettcher, T. B. & Jayappa, H. Comparison of antibody production , lymphocyte stimulation , and protection induced by four commercial Mycoplasma hyopneumoniae bacterins. *Swine Heal. Prod.* **6,** 107–112 (1998).
11. Djordjevic, S. P. *et al.* Serum and mucosal antibody responses and protection in pigs vaccinated against Mycoplasma hyopneumoniae with vaccines containing a denatured membrane antigen pool and adjuvant. *Aust. Vet. J.* **75,** 504–511 (1997).
12. Vranckx, K. *et al.* Vaccination reduces macrophage infiltration in bronchus-associated lymphoid tissue in pigs infected with a highly virulent Mycoplasma hyopneumoniae strain. *BMC Vet. Res.* **8,** 24 (2012).
13. Simionatto, S., Marchioro, S. B., Maes, D. & Dellagostin, O. A. Mycoplasma hyopneumoniae: From disease to vaccine development. *Vet. Microbiol.* **165,** 234–242 (2013).
14. Razin, S., Yogev, D. & Naot, Y. Molecular biology and pathogenicity of mycoplasmas. *Microbiol. Mol. Biol. Rev.* **62,** 1094–156 (1998).
15. Minion, F. C. *et al.* The genome sequence of Mycoplasma hyopneumoniae strain 232, the agent of swine mycoplasmosis. *J. Bacteriol.* **186,** 7123–7133 (2004).
16. Vasconcelos, A. T. R. *et al.* Swine and Poultry Pathogens : the Complete Genome Sequences of Two Strains of Mycoplasma hyopneumoniae and a Strain of Mycoplasma synoviae †. **187,** 5568–5577 (2005).
17. Liu, W. *et al.* Comparative genomic analyses of Mycoplasma hyopneumoniae pathogenic 168 strain and its high-passaged attenuated strain. *BMC Genomics* **14,** 80 (2013).
18. Zielinski, G. C. & Ross, R. F. Effect of growth in cell cultures and strain on virulence of Mycoplasma hyopneumoniae for swine. *Am. J. Vet. Res.* **51,** 344—348 (1990).
19. Zielinski, G. C., Young, T., Ross, R. F. & Rosenbusch, R. F. Adherence of Mycoplasma hyopneumoniae to cell monolayers. *Am. J. Vet. Res.* **51,** 339—343 (1990).
20. Yogev, D., Browning, G. F. & Wise, K. S. in *Molecular Biology and Pathogenicity of Mycoplasmas* (eds. Razin, S. & Herrmann, R.) 417–443 (K1uwer Academic/Plenum Publishers, 2002).

21. Blanchard, B. *et al.* Electron microscopic observation of the respiratory tract of SPF piglets inoculated with Mycoplasma hyopneumoniae. *Vet. Microbiol.* **30,** 329–341 (1992).
22. Mattsson, J. G., Bergström, K., Wallgren, P., Johansson, K. E. & Wallgren, P. E. R. Detection of Mycoplasma hyopneumoniae in nose swabs from pigs by in vitro amplification of the 16S rRNA gene. *J. Clin. Microbiol.* **33,** 893–897 (1995).
23. Zhang, Q., Young, T. F. & Ross, R. F. Microtiter plate adherence assay and receptor analogs for Mycoplasma hyopneumoniae. *Infect. Immun.* **62,** 1616–1622 (1994).
24. Zhang, Q., Young, T. F., Ross, R. F., Zhang, Q. & Young, T. F. Identification and characterization of a Mycoplasma hyopneumoniae adhesin. *Infect. Immun.* **63,** 1013–1019 (1995).
25. Jenkins, C. *et al.* Two domains within the Mycoplasma hyopneumoniae cilium adhesin bind heparin. *Infect. Immun.* **74,** 481–487 (2006).
26. Hsu, T. & Chris Minion, F. Identification of the cilium binding epitope of the Mycoplasma hyopneumoniae P97 adhesin. *Infect. Immun.* **66,** 4762–4766 (1998).
27. Seymour, L. M. *et al.* Mhp182 (P102) binds fibronectin and contributes to the recruitment of plasmin(ogen) to the Mycoplasma hyopneumoniae cell surface. *Cell. Microbiol.* **14,** 81–94 (2012).
28. Seymour, L. M. *et al.* Mhp107 is a member of the multifunctional adhesin family of Mycoplasma hyopneumoniae. *J. Biol. Chem.* **286,** 10097–104 (2011).
29. Wilton, J. *et al.* Mhp493 (P216) is a proteolytically processed, cilium and heparin binding protein of Mycoplasma hyopneumoniae. *Mol. Microbiol.* **71,** 566–82 (2009).
30. Deutscher, A. T. *et al.* Repeat regions R1 and R2 in the P97 paralogue Mhp271 of Mycoplasma hyopneumoniae bind heparin, fibronectin and porcine cilia. *Mol. Microbiol.* **78,** 444–458 (2010).
31. Deutscher, A. T. *et al.* Mycoplasma hyopneumoniae Surface proteins Mhp385 and Mhp384 bind host cilia and glycosaminoglycans and are endoproteolytically processed by proteases that recognize different cleavage motifs. *J. Proteome Res.* **11,** 1924–1936 (2012).
32. Bogema, D. R. *et al.* Sequence TTKF ↓ QE defines the site of proteolytic cleavage in Mhp683 protein, a novel glycosaminoglycan and cilium adhesin of Mycoplasma hyopneumoniae. *J. Biol. Chem.* **286,** 41217–41229 (2011).
33. Raymond, B. B. A. *et al.* P159 from Mycoplasma hyopneumoniae binds porcine cilia and heparin and is cleaved in a manner akin to ectodomain shedding. *J. Proteome Res.* **12,** 5891–5903 (2013).
34. Robinson, M. W. M. W. M. W. *et al.* MHJ_0125 is an M42 glutamyl aminopeptidase that moonlights as a multifunctional adhesin on the surface of Mycoplasma hyopneumoniae. *Open Biol.* **3,** 130017 (2013).
35. DeBey, M. C. & Ross, R. F. Ciliostasis and loss of cilia induced by Mycoplasma hyopneumoniae in porcine tracheal organ cultures. *Infect. Immun.* **62,** 5312–5318 (1994).
36. Vilei, E. M. & Frey, J. Genetic and biochemical characterization of glycerol uptake in mycoplasma mycoides subsp. mycoides SC: its impact on H(2)O(2) production and virulence. *Clin. Diagn. Lab. Immunol.* **8,** 85–92 (2001).
37. Bogema, D. R. *et al.* Characterization of Cleavage Events in the Multifunctional Cilium Adhesin Mhp684 (P146) Reveals a Mechanism by Which Mycoplasma hyopneumoniae Regulates Surface Topography. *MBio* **3,** 1–11 (2012).
38. Lähteenmäki, K., Edelman, S. & Korhonen, T. K. Bacterial metastasis: the host plasminogen system in bacterial invasion. *Trends Microbiol.* **13,** 79–85 (2005).
39. Wise, K. I. M. S. & Kim, M. F. Major membrane surface proteins of Mycoplasma hyopneumoniae selectively modified by covalently bound lipid. *J. Bacteriol.* **169,** 5546–5555 (1987).
40. Tajima, M. & Yagihashi, T. Interaction of Mycoplasma hyopneumoniae with the Porcine Respiratory Epithelium as Observed by Electron Microscopy. *Infect. Immun.* **37,** 1162–1169 (1982).
41. Meyns, T. *et al.* Interactions of highly and low virulent Mycoplasma hyopneumoniae isolates

with the respiratory tract of pigs. *Vet. Microbiol.* **120,** 87–95 (2007).
42. Choi, C. *et al.* Expression of inflammatory cytokines in pigs experimentally infected with Mycoplasma hyopneumoniae. *J. Comp. Pathol.* **134,** 40–46 (2006).
43. Ramírez, A. S. *et al.* A semi-defined medium without serum for small ruminant mycoplasmas. *Vet. J.* **178,** 149–52 (2008).
44. Rodríguez, F., Ramírez, G. A., Sarradell, J., Andrada, M. & Lorenzo, H. Immunohistochemical labelling of cytokines in lung lesions of pigs naturally infected with Mycoplasma hyopneumoniae. *J. Comp. Pathol.* **130,** 306–312 (2004).
45. Schmidt, J. A., Browning, G. F. & Markham, P. F. Mycoplasma hyopneumoniae mhp379 is a Ca2+-dependent, sugar-nonspecific exonuclease exposed on the cell surface. *J. Bacteriol.* **189,** 3414–3424 (2007).
46. Staats, C. C., Boldo, J., Broetto, L., Vainstein, M. & Schrank, A. Comparative genome analysis of proteases, oligopeptide uptake and secretion systems in Mycoplasma spp. *Genet. Mol. Biol.* **30,** 225–229 (2007).
47. Schmidt, J. A., Browning, G. F. & Markham, P. F. Mycoplasma hyopneumoniae p65 surface lipoprotein is a lipolytic enzyme with a preference for shorter-chain fatty acids. *J. Bacteriol.* **186,** 5790–5798 (2004).
48. Rottem, S. Interaction of mycoplasmas with host cells. *Physiol. Rev.* **83,** 417–432 (2003).
49. Frandson, R. D., Lee Wilke, W. & Fails, A. D. *Anatomy and Physiology of Farm Animals*. (John Wiley & Sons, 2013).
50. Staff of Fort Dodge, Baulkham Hills, N. S. W. Respiratory disease in pigs. Available at: http://www.veterinaria.org/revistas/vetenfinf/vet_enf_inf_tripod/porcinos/Respiratorydiseaseinpigs.htm. (Accessed: 23rd June 2017)
51. Siqueira, F. M., De Souto Weber, S., Cattani, A. M. & Schrank, I. S. Genome organization in Mycoplasma hyopneumoniae: Identification of promoter-like sequences. *Mol. Biol. Rep.* **41,** 5395–5402 (2014).
52. Weber, S. D. S., Sant'Anna, F. H. & Schrank, I. S. Unveiling Mycoplasma hyopneumoniae promoters: sequence definition and genomic distribution. *DNA Res.* **19,** 103–15 (2012).
53. Fritsch, T. E., Siqueira, F. M. & Schrank, I. S. Intrinsic terminators in Mycoplasma hyopneumoniae transcription. *BMC Genomics* **16,** 273 (2015).
54. Siqueira, F. M. *et al.* Unravelling the transcriptome profile of the Swine respiratory tract mycoplasmas. *PLoS One* **9,** e110327 (2014).
55. Güell, M. *et al.* Transcriptome complexity in a genome-reduced bacterium. *Science (80-.).* **326,** 1268–1271 (2009).
56. Junier, I., Unal, E. B., Yus, E., Lloréns-Rico, V. & Serrano, L. Insights into the Mechanisms of Basal Coordination of Transcription Using a Genome-Reduced Bacterium. *Cell Syst.* **2,** 391–401 (2016).
57. Cattani, A. M., Siqueira, F. M., Muniz Guedes, R. L. & Schrank, I. S. Repetitive elements in mycoplasma hyopneumoniae transcriptional regulation. *PLoS One* **11,** e0168626 (2016).
58. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30,** 1236–1240 (2014).
59. Koehorst, J. J. *et al.* Comparison of 432 Pseudomonas strains through integration of genomic, functional, metabolic and expression data. *Sci. Rep.* **6,** 38699 (2016).
60. Orth, J. D., Thiele, I. & Palsson, B. Ø. What is flux balance analysis? *Nat. Biotechnol.* **28,** 245–248 (2010).
61. Varma, A. & Palsson, B. O. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110. *Appl. Environ. Microbiol.* **60,** 3724–3731 (1994).
62. Henry, C. S. *et al.* High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat. Biotechnol.* **28,** 977–982 (2010).
63. Karp, P. D. *et al.* Pathway tools version 19.0 update: Software for pathway/genome informatics and systems biology. *Brief. Bioinform.* **17,** 877–890 (2016).
64. Thiele, I. & Palsson, B. Ø. A protocol for generating a high-quality genome-scale metabolic

reconstruction. *Nat. Protoc.* **5,** 93–121 (2010).
65. Oberhardt, M. A., Palsson, B. Ø. & Papin, J. A. Applications of genome-scale metabolic reconstructions. *Mol. Syst. Biol.* **5,** 1–15 (2009).
66. Lee, S. Y. & Kim, H. U. Systems strategies for developing industrial microbial strains. *Nat. Biotechnol.* **33,** 1061–1072 (2015).
67. Wodke, J. a H. *et al.* Dissecting the energy metabolism in Mycoplasma pneumoniae through genome-scale metabolic modeling. *Mol. Syst. Biol.* **9,** 653 (2013).
68. Bautista, E. J. *et al.* Semi-automated Curation of Metabolic Models via Flux Balance Analysis: A Case Study with Mycoplasma gallisepticum. *PLoS Comput. Biol.* **9,** e1003208 (2013).
69. Suthers, P. F. *et al.* A genome-scale metabolic reconstruction of Mycoplasma genitalium, iPS189. *PLoS Comput. Biol.* **5,** 1–14 (2009).
70. Sirand-Pugnet, P. *et al.* Being pathogenic, plastic, and sexual while living with a nearly minimal bacterial genome. *PLoS Genet.* **3,** 744–758 (2007).
71. Sirand-Pugnet, P., Citti, C., Barré, A. & Blanchard, A. Evolution of mollicutes: down a bumpy road with twists and turns. *Res. Microbiol.* **158,** 754–766 (2007).
72. Citti, C. & Blanchard, A. Mycoplasmas and their host: Emerging and re-emerging minimal pathogens. *Trends Microbiol.* **21,** 196–203 (2013).
73. Glass, J. I. *et al.* Essential genes of a minimal bacterium. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 425–430 (2006).
74. Kuhner, S. *et al.* Proteome organization in a genome-reduced bacterium. *Science (80-.).* **326,** 1235–40 (2009).
75. Yus, E. *et al.* Impact of genome reduction on bacterial metabolism and its regulation. *Science* **326,** 1263–8 (2009).
76. WHO. *Global Vaccine Action Plan 2011-2020*. (2013).
77. Josefsberg, J. O. & Buckland, B. Vaccine process technology. *Biotechnol. Bioeng.* **109,** 1443–1460 (2012).
78. Heldens, J. G. M. *et al.* Veterinary vaccine development from an industrial perspective. *Vet. J.* **178,** 7–20 (2008).
79. Grand View Research. *Vaccine Market Analysis By Type (Live Attenuated Vaccines, Inactivated Vaccines, Subunit Vaccines, Toxoid Vaccines, Conjugate Vaccines, DNA Vaccines), By Application (Infectious Diseases, Cancer, Autism, Allergy) And Segment Forecasts To 2024*. (2016).
80. Way, J. C., Collins, J. J., Keasling, J. D. & Silver, P. A. Integrating biological redesign: Where synthetic biology came from and where it needs to go. *Cell* **157,** 151–161 (2014).
81. Gustavsson, M. & Lee, S. Y. Prospects of microbial cell factories developed through systems metabolic engineering. *Microb. Biotechnol.* **9,** 610–617 (2016).
82. Jullesson, D., David, F., Pfleger, B. & Nielsen, J. Impact of synthetic biology and metabolic engineering on industrial production of fine chemicals. *Biotechnol. Adv.* **33,** 1395–1402 (2015).
83. Junker, B. *et al.* Design-for-Six-Sigma for Development of a Bioprocess Quality-by-Design Framework. *PDA J. Pharm. Sci. Technol.* **65,** 254–286 (2011).
84. Teng, S.-H. & Ho, S.-Y. Failure mode and effects analysis An integrated approach for product design and process control. *Int. J. Qual. Reliab. Manag.* **13,** 8–26 (2006).
85. Nielsen, J. & Keasling, J. D. Engineering Cellular Metabolism. *Cell* **164,** 1185–1197 (2016).
86. Hutchison, C. A. *et al.* Design and synthesis of a minimal bacterial genome. *Science (80-.).* **351,** aad6253-1-11 (2016).
87. Rogers, J. K. & Church, G. M. Multiplexed Engineering in Biology. *Trends Biotechnol.* **34,** 198–206 (2016).
88. Pronk, J. T. *et al.* How to set up collaborations between academia and industrial biotech companies. *Nat. Biotechnol.* **33,** 237–240 (2015).
89. Bartell, J. A., Yen, P., Varga, J. J., Goldberg, J. B. & Papin, J. A. Comparative metabolic systems analysis of pathogenic Burkholderia. *J. Bacteriol.* **196,** 210–226 (2014).

90. Bartell, J. A. *et al.* Reconstruction of the metabolic network of Pseudomonas aeruginosa to interrogate virulence factor synthesis. *Nat. Commun.* **8,** 14631 (2017).
91. Minato, Y., Fassio, S. R., Wolfe, A. J. & Häse, C. C. Central metabolism controls transcription of a virulence gene regulator in Vibrio cholerae. *Microbiology* **159,** 792–802 (2013).
92. Bouillaut, L., Dubois, T., Sonenshein, A. L. & Dupuy, B. Integration of metabolism and virulence in Clostridium difficile. *Res. Microbiol.* **166,** 375–383 (2015).
93. Brown, A. J. P., Brown, G. D., Netea, M. G. & Gow, N. A. R. Metabolism impacts upon candida immunogenicity and pathogenicity at multiple levels. *Trends Microbiol.* **22,** 614–622 (2014).
94. Eisenreich, W., Dandekar, T., Heesemann, J. & Goebel, W. Carbon metabolism of intracellular bacterial pathogens and possible links to virulence. *Nat. Rev. Microbiol.* **8,** 401–412 (2010).
95. Burgard, A. P., Pharkya, P. & Maranas, C. D. OptKnock: A Bilevel Programming Framework for Identifying Gene Knockout Strategies for Microbial Strain Optimization. *Biotechnol. Bioeng.* **84,** 647–657 (2003).
96. Cheng, J. K. & Alper, H. S. The genome editing toolbox: A spectrum of approaches for targeted modification. *Curr. Opin. Biotechnol.* **30,** 87–94 (2014).
97. Esvelt, K. M. & Wang, H. H. Genome-scale engineering for systems and synthetic biology. *Mol. Syst. Biol.* **9,** 641 (2013).
98. Tsai, S. Q. *et al.* GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat Biotechnol* **33,** 187–197 (2015).
99. Lin, Y. *et al.* CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences. *Nucleic Acids Res.* **42,** 7473–7485 (2014).
100. Lluch-Senar, M. *et al.* Defining a minimal cell: essentiality of small ORFs and ncRNAs in a genome-reduced bacterium. *Mol. Syst. Biol.* **11,** 780 (2015).
101. Turner, K. H., Wessel, A. K., Palmer, G. C., Murray, J. L. & Whiteley, M. Essential genome of *Pseudomonas aeruginosa* in cystic fibrosis sputum. *Proc. Natl. Acad. Sci.* **112,** 4110–4115 (2015).
102. Liu, X. *et al.* High-throughput CRISPRi phenotyping in Streptococcus pneumoniae identifies new essential genes involved in cell wall synthesis and competence development. *Mol. Syst. Biol.* **13,** 1–18 (2017).
103. van Opijnen, T. & Camilli, A. Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms. *Nat. Rev. Microbiol.* **11,** 435–442 (2013).
104. Schoen, C., Kischkies, L., Elias, J. & Ampattu, B. J. Metabolism and virulence in Neisseria meningitidis. *Front. Cell. Infect. Microbiol.* **4,** 1–16 (2014).
105. Galen, J. E. & Curtiss, R. The delicate balance in genetically engineering live vaccines. *Vaccine* **32,** 4376–4385 (2014).
106. Rathore, A. S. Quality by Design (QbD)-Based Process Development for Purification of a Biotherapeutic. *Trends Biotechnol.* **34,** 358–370 (2016).
107. Haas, J. *et al.* Implementation of QbD for the development of a vaccine candidate. *Vaccine* **32,** 2927–2930 (2014).
108. Van De Waterbeemd, B. *et al.* Gene-expression-based quality scores indicate optimal harvest point in Bordetella pertussis cultivation for vaccine production. *Biotechnol. Bioeng.* **103,** 900–908 (2009).
109. Licona-Cassani, C. *et al.* Tetanus toxin production is triggered by the transition from amino acid consumption to peptides. *Anaerobe* **41,** 113–124 (2016).
110. Mercier, S. M., Diepenbroek, B., Wijffels, R. H. & Streefland, M. Multivariate PAT solutions for biopharmaceutical cultivation: Current progress and limitations. *Trends Biotechnol.* **32,** 329–336 (2014).
111. Glassey, J. *et al.* Process analytical technology (PAT) for biopharmaceuticals. *Biotechnol. J.* **6,** 369–377 (2011).
112. Rodwell, A. W. A Defined Medium for Mycoplasma Strain Y. *J. gen. Microbiol.* **58,** 39–47

(1969).
113. Tourtelotte, M. E., Morowitz, H. J. & Kasimer, P. Defined Medium for Mycoplasma Laidlawii. *J. Bacteriol.* **88,** 11–15 (1964).
114. Schilling, C. H. & Palsson, B. O. Assessment of the metabolic capabilities of Haemophilus influenzae Rd through a genome-scale pathway analysis. *J. Theor. Biol.* **203,** 249–283 (2000).
115. Juty, N. *et al.* BioModels: Content, features, functionality, and use. *CPT Pharmacometrics Syst. Pharmacol.* **4,** 55–68 (2015).
116. Ferrarini, M. G. *et al.* Insights on the virulence of swine respiratory tract mycoplasmas through genome-scale metabolic modeling. *BMC Genomics* **17,** 353 (2016).
117. Friis, N. F. Some recommendations concerning primary isolation of Mycoplasma suipneumoniae and Mycoplasma flocculare a survey. *Nord. Vet. Med.* **27,** 337—339 (1975).
118. Rosengarten, R., Fischer, M., Kirchhoff, H., Kerlen, G. & Seack, K. Transport of Erythrocytes by Gliding Cells of Mycoplasma mobile 163K. *Curr. Microbiol.* **16,** 253–257 (1988).
119. Piccolo, S. R. *et al.* A single-sample microarray normalization method to facilitate personalized-medicine workflows. *Genomics* **100,** 337–344 (2012).
120. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11,** 119 (2010).
121. Karp, P. D., Latendresse, M. & Caspi, R. The pathway tools pathway prediction algorithm. *Stand. Genomic Sci.* **5,** 424–429 (2011).
122. Pollack, J. D. The necessity of combining genomic and enzymatic data to infer metabolic function and pathways in the smallest bacteria: amino acid, purine and pyrimidine metabolism in Mollicutes. *Front. Biosci.* **7,** 1762–1781 (2002).
123. Ebrahim, A., Lerman, J. A., Palsson, B. O. & Hyduke, D. R. COBRApy: COnstraints-Based Reconstruction and Analysis for Python. *BMC Syst. Biol.* **7,** 74 (2013).
124. Schellenberger, J. *et al.* Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat. Protoc.* **6,** 1290–1307 (2011).
125. Bornstein, B. J., Keating, S. M., Jouraku, A. & Hucka, M. LibSBML: An API library for SBML. *Bioinformatics* **24,** 880–881 (2008).
126. Miles, R. J., Beezer, A. E. & Lee, D. H. Kinetics of utilization of organic substrates by Mycoplasma mycoides subsp. mycoides in a salts solution: a flow-microcalorimetric study. *J. Gen. Microbiol.* **131,** 1845–1852 (1985).
127. Razin, S., Argaman, M. & Avigan, J. Chemical Composition of Mycoplasma Cells and Membranes. *J. Gen. Microbiol.* **33,** 477–487 (1963).
128. Mahadevan, R. & Schilling, C. H. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab. Eng.* **5,** 264–276 (2003).
129. Chelliah, V. *et al.* BioModels: Ten-year anniversary. *Nucleic Acids Res.* **43,** D542–D548 (2015).
130. Hjersted, J. L. & Henson, M. A. Optimization of fed-batch Saccharomyces cerevisiae fermentation using dynamic flux balance models. *Biotechnol. Prog.* **22,** 1239–1248 (2006).
131. Koehorst, J. J., Saccenti, E., Schaap, P. J., Martins dos Santos, V. A. P. & Suarez-Diez, M. Protein domain architectures provide a fast, efficient and scalable alternative to sequence-based methods for comparative functional genomics. *F1000Research* **5,** 1987 (2016).
132. Piccolo, S. R., Withers, M. R., Francis, O. E., Bild, A. H. & Johnson, W. E. Multiplatform single-sample estimates of transcriptional activation. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 17778–17783 (2013).
133. Maglennon, G. A. *et al.* Transposon mutagenesis in Mycoplasma hyopneumoniae using a novel mariner-based system for generating random mutations. *Vet. Res.* **44,** 1–11 (2013).
134. Miles, R. J., Wadher, B. J., Henderson, C. L., Mohan, K. & Henderson, L. Increased growth yields of Mycoplasma spp. in the presence of pyruvate. *Lett. Appl. Microbiol.* **7,** 149–151 (1988).
135. de Crécy-Lagard, V., El Yacoubi, B., de la Garza, R. D., Noiriel, A. & Hanson, A. D.

Comparative genomics of bacterial and plant folate synthesis and salvage: predictions and validations. *BMC Genomics* **8,** 245 (2007).
136. Newton, D. T., Creuzenet, C. & Mangroo, D. Formylation Is Not Essential for Initiation of Protein Synthesis in All Eubacteria. *J. Biol. Chem.* **274,** 22143–22146 (1999).
137. Gil, R., Silva, F. J., Peretó, J. & Moya, A. Determination of the Core of a Minimal Bacterial Gene Set †. *Microbiol. Mol. Biol. Rev.* **68,** 518–537 (2004).
138. Rosengarten, R. *et al.* Host-pathogen interactions in mycoplasma pathogenesis: virulence and survival strategies of minimalist prokaryotes. *Int. J. Med. Microbiol.* **290,** 15–25 (2000).
139. Browning, G. F., Marenda, M. S., Noormohammadi, A. H. & Markham, P. F. The central role of lipoproteins in the pathogenesis of mycoplasmoses. *Vet. Microbiol.* **153,** 44–50 (2011).
140. Großhennig, S. *et al.* Hydrogen Sulfide is a Novel Potential Virulence Factor of Mycoplasma pneumoniae: Characterization of the Unusual Cysteine Desulfurase/Desulfhydrase HapE. *Mol. Microbiol.* **100,** 42–54 (2015).
141. Gibson, D. G. *et al.* Creation of a Bacterial Cell Controlled by a Chemically Synthesized Genome. *Science (80-.).* **329,** 52–56 (2010).
142. Rouli, L., Merhej, V., Fournier, P.-E. & Raoult, D. The bacterial pangenome as a new tool for analyzing pathogenic bacteria. *New Microbes New Infect.* **7,** 72–85 (2015).
143. Liu, W. *et al.* Comparative genomics of Mycoplasma: analysis of conserved essential genes and diversity of the pan-genome. *PLoS One* **7,** e35698 (2012).
144. Guimaraes, A. M. S., Santos, A. P., Do Nascimento, N. C., Timenetsky, J. & Messick, J. B. Comparative genomics and phylogenomics of hemotrophic mycoplasmas. *PLoS One* **9,** (2014).
145. Saccenti, E., Nieuwenhuijse, D., Koehorst, J. J., Dos Santos, V. A. P. M. & Schaap, P. J. Assessing the metabolic diversity of streptococcus from a protein domain point of view. *PLoS One* **10,** 1–20 (2015).
146. Kuznetsov, V., Pickalov, V. & Kanapin, A. in *Bioinformatics of Genome Regulation and Structure II* (eds. Kolchanov, N., Hofestaedt, R. & Milanesi, L.) 329–341 (Springer, 2006). doi:10.1007/0-387-29455-4_32
147. Weisburg, W. G. *et al.* A phylogenetic analysis of the mycoplasmas: basis for their classification. *J. Bacteriol.* **171,** 6455–6467 (1989).
148. Temple Lang, D. RCurl: General Network (HTTP/FTP/...) Client Interface for R. (2015).
149. Van Hage, W. R. SPARQL: SPARQL client. (2013).
150. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res.* **41,** 590–596 (2013).
151. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* **20,** 289–290 (2004).
152. Galili, T. dendextend: An R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics* **31,** 3718–3720 (2015).
153. Snipen, L., Almøy, T. & Ussery, D. W. Microbial comparative pan-genomics using binomial mixture models. *BMC Genomics* **10,** 385 (2009).
154. Wolf, Y. I. & Koonin, E. V. A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes. *Genome Biol. Evol.* **4,** 1286–1294 (2012).
155. Ekseth, O. K., Kuiper, M. & Mironov, V. OrthAgogue: An agile tool for the rapid prediction of orthology relations. *Bioinformatics* **30,** 734–736 (2014).
156. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30,** 1575–84 (2002).
157. Breiman, L. Random forests. *Mach. Learn.* **45,** 5–32 (2001).
158. Venables, W. N. & Ripley, B. D. in *Modern Applied Statistics with S* (eds. Chambers, J., Eddy, W., Härdle, W., Sheather, S. & Tierney, L.) 271–300 (Springer New York, 2002). doi:10.1007/978-0-387-21706-2_10
159. Liaw, A. & Wiener, M. Classification and Regression by randomForest. *R News* **2,** 18–22 (2002).

160. Tripp, H. J. *et al.* Toward a standard in structural genome annotation for prokaryotes. *Stand. Genomic Sci.* **10,** 45 (2015).
161. Guimaraes, A. M. S. *et al.* Complete genome sequence of Mycoplasma suis and insights into its biology and adaption to an erythrocyte niche. *PLoS One* **6,** e19574 (2011).
162. Do Nascimento, N. C., Santos, A. P., Guimaraes, A. M. S., Sanmiguel, P. J. & Messick, J. B. Mycoplasma haemocanis - The canine hemoplasma and its feline counterpart in the genomic era. *Vet. Res.* **43,** 66 (2012).
163. Hunter, S. *et al.* InterPro in 2011: New developments in the family and domain prediction database. *Nucleic Acids Res.* **40,** 1–7 (2012).
164. Santos, A. P. *et al.* Genome of Mycoplasma haemofelis, unraveling its strategies for survival and persistence. *Vet. Res.* **42,** 102 (2011).
165. Mcmenamy, R. H., Lund, C. C. & Oncley, J. L. Unbound amino acid concentrations in human blood plasmas. *J. Clin. Invest.* **1,** 1672–1679 (1957).
166. Díaz-Magaña, A. *et al.* Short-chain chromate ion transporter proteins from Bacillus subtilis confer chromate resistance in Escherichia coli. *J. Bacteriol.* **191,** 5441–5445 (2009).
167. Pollack, J. D., Williams, M. V & McElhaney, R. N. 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.* **23,** 269–354 (1997).
168. Eberl, M., Hintz, M., Jamba, Z., Beck, E. & Jomaa, H. Mycoplasma penetrans Is Capable of Activating Vγ9/Vδ2 T Cells While Other Human Pathogenic Mycoplasmas Fail To Do So. *Infect. Immun.* **72,** 4881–4883 (2004).
169. Ben-Menachem, G., Himmelreich, R., Herrmann, R., Aharonowitz, Y. & Rottem, S. The thioredoxin reductase system of mycoplasmas. *Microbiology* **143,** 1933–1940 (1997).
170. Chen, X. & Ishwaran, H. Random forests for genomic data analysis. *Genomics* **99,** 323–329 (2012).
171. Nouvel, L. X. *et al.* Comparative genomic and proteomic analyses of two Mycoplasma agalactiae strains: clues to the macro- and micro-events that are shaping mycoplasma diversity. *BMC Genomics* **11,** 86 (2010).
172. Dandekar, T. *et al.* Re-annotating the Mycoplasma pneumoniae genome sequence: adding value, function and reading frames. *Nucleic Acids Res.* **28,** 3278–3288 (2000).
173. Maier, T. *et al.* Large-scale metabolome analysis and quantitative integration with genomics and proteomics data in Mycoplasma pneumoniae. *Mol. Biosyst.* **9,** 1743–1755 (2013).
174. Krause, D. C. Mycoplasma pneumoniae cytadherence: Unravelling the tie that binds. *Mol. Microbiol.* **20,** 247–253 (1996).
175. Pflaum, K., Tulman, E. R., Beaudet, J., Liao, X. & Geary, S. J. Global changes in Mycoplasma gallisepticum phase-variable lipoprotein gene vlhA expression during in vivo infection of the natural chicken host. *Infect. Immun.* **84,** 351–355 (2015).
176. Huang, S., Li, J. Y., Wu, J., Meng, L. & Shou, C. C. Mycoplasma infections and different human carcinomas. *World J. Gastroenterol.* **7,** 266–269 (2001).
177. Pitcher, D. G. & Nicholas, R. A. J. Mycoplasma host specificity: Fact or fiction? *Vet. J.* **170,** 300–306 (2005).
178. Tu, A. T., Voelker, L. L., Shen, X. & Dybvig, K. Complete nucleotide sequence of the mycoplasma virus P1 genome. *Plasmid* **45,** 122–6 (2001).
179. Galperin, M. Y. & Koonin, E. V. 'Conserved hypothetical' proteins: Prioritization of targets for experimental study. *Nucleic Acids Res.* **32,** 5452–5463 (2004).
180. Maes, D., Verdonck, M., Deluyker, H. & de Kruif, A. Enzootic pneumonia in pigs. *Vet. Q.* **18,** 104–109 (1996).
181. Bin, L. *et al.* Transcription analysis of the porcine alveolar macrophage response to Mycoplasma hyopneumoniae. *PLoS One* **9,** e101968 (2014).
182. Djordjevic, S. P. *et al.* Proteolytic Processing of the Mycoplasma hyopneumoniae Cilium Adhesin. **72,** (2004).
183. Seymour, L. M. *et al.* A processed multidomain mycoplasma hyopneumoniae adhesin binds

fibronectin, plasminogen, and swine respiratory cilia. *J. Biol. Chem.* **285,** 33971–33978 (2010).
184. Adams, C., Pitzer, J. & Minion, F. C. In Vivo Expression Analysis of the P97 and P102 Paralog Families of Mycoplasma hyopneumoniae. *Infect. Immun.* **73,** 7784–7787 (2005).
185. Reolon, L. A., Martello, C. L., Schrank, I. S. & Ferreira, H. B. Survey of surface proteins from the pathogenic mycoplasma hyopneumoniae strain 7448 using a biotin cell surface labeling approach. *PLoS One* **9,** (2014).
186. Ferreira, H. B. & Castro, L. A. De. A preliminary survey of M . hyopneumoniae virulence factors based on comparative genomic analysis. **255,** 245–255 (2007).
187. Brennan, P. C. & Feinstein, R. N. Relationship of Hydrogen Peroxide Production by Mycoplasma pulmonis to Virulence for Catalase- deficient Mice. *J. Bacteriol.* **98,** 1036–1040 (1969).
188. Schmidl, S. R. *et al.* A trigger enzyme in Mycoplasma pneumoniae: impact of the glycerophosphodiesterase GlpQ on virulence and gene expression. *PLoS Pathog.* **7,** e1002263 (2011).
189. Madsen, M. L., Puttamreddy, S., Thacker, E. L., Carruthers, M. D. & Minion, F. C. Transcriptome changes in Mycoplasma hyopneumoniae during infection. *Infect. Immun.* **76,** 658–663 (2008).
190. Croucher, N. J. & Thomson, N. R. Studying bacterial transcriptomes using RNA-seq. *Curr. Opin. Microbiol.* **13,** 619–624 (2010).
191. Boisvert, S., Laviolette, F. & Corbeil, J. Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *J. Comput. Biol.* **17,** 1519–1533 (2010).
192. Camacho, C. *et al.* BLAST plus: architecture and applications. *BMC Bioinformatics* **10,** 1 (2009).
193. Mengeling, W. L., Lager, K. M. & Vorwald, A. C. Diagnosis of Porcine Reproductive and Respiratory Syndrome. *J. Vet. Diagnostic Investig.* **7,** 3–16 (1995).
194. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14,** R36 (2013).
195. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
196. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
197. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2009).
198. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **40,** 4288–4297 (2012).
199. Rienksma, R. A. *et al.* Comprehensive insights into transcriptional adaptation of intracellular mycobacteria by microbe-enriched dual RNA sequencing. *BMC Genomics* **16,** 34 (2015).
200. Robinson, M. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11,** R25 (2010).
201. Robinson, M. D. & Smyth, G. K. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics* **9,** 321–332 (2008).
202. Robinson, M. D. & Smyth, G. K. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* **23,** 2881–2887 (2007).
203. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing. *J R Stat. Soc B* **57,** 289–300 (1995).
204. Robbe-Saule, M., Babonneau, J., Sismeiro, O., Marsollier, L. & Marion, E. An Optimized Method for Extracting Bacterial RNA from Mouse Skin Tissue Colonized by Mycobacterium ulcerans. *Front. Microbiol.* **8,** 1–10 (2017).
205. Westermann, A. J., Gorski, S. a. & Vogel, J. Dual RNA-seq of pathogen and host. *Nat. Rev. Microbiol.* **10,** 618–630 (2012).

206. Béven, L. *et al.* Specific Evolution of F1-Like ATPases in Mycoplasmas. *PLoS One* **7,** e38793 (2012).
207. Paddenberg, R., Weber, a, Wulf, S. & Mannherz, H. G. Mycoplasma nucleases able to induce internucleosomal DNA degradation in cultured cells possess many characteristics of eukaryotic apoptotic nucleases. *Cell Death Differ.* **5,** 517–528 (1998).
208. Bendjennat, M., Blanchard, A., Loutfi, M., Montagnier, L. & Bahraoui, E. Role of Mycoplasma penetrans endonuclease P40 as a potential pathogenic determinant. *Infect. Immun.* **67,** 4456–4462 (1999).
209. Cacciotto, C. *et al.* Mycoplasma lipoproteins are major determinants of neutrophil extracellular trap formation. *Cell. Microbiol.* **18,** 1751–1762 (2016).
210. Wang, C. *et al.* The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. *Nat. Biotechnol.* **32,** 926–932 (2014).
211. Papenfort, K. & Vogel, J. Regulatory RNA in bacterial pathogens. *Cell Host Microbe* **8,** 116–127 (2010).
212. Vogel, J. A rough guide to the non-coding RNA world of Salmonella. *Mol. Microbiol.* **71,** 1–11 (2009).
213. Lee, E. J. & Groisman, E. A. An antisense RNA that governs the expression kinetics of a multifunctional virulence gene. *Mol. Microbiol.* **76,** 1020–1033 (2010).
214. Altuvia, S. Identification of bacterial small non-coding RNAs: experimental approaches. *Curr. Opin. Microbiol.* **10,** 257–261 (2007).
215. Thomason, M. K. & Storz, G. Bacterial Antisense RNAs: How Many Are There, and What Are They Doing? *Annu. Rev. Genet.* **44,** 167–188 (2010).
216. Lybecker, M., Bilusic, I. & Raghavan, R. Pervasive transcription: detecting functional RNAs in bacteria. *Transcription* **5,** 1–5 (2014).
217. Wade, J. T. & Grainger, D. C. Pervasive transcription: illuminating the dark matter of bacterial transcriptomes. *Nat. Rev. Microbiol.* **12,** 647–653 (2014).
218. Raghavan, R., Sloan, D. B. & Ochman, H. Antisense Transcription Is Pervasive but Rarely Conserved in Enteric Bacteria. *MBio* **3,** (2012).
219. Gottesman, S. & Storz, G. Bacterial small RNA regulators: Versatile roles and rapidly evolving variations. *Cold Spring Harb. Perspect. Biol.* **3,** 1–16 (2011).
220. Storz, G., Vogel, J. & Wassarman, K. M. Regulation by Small RNAs in Bacteria: Expanding Frontiers. *Molecular Cell* **43,** 880–891 (2011).
221. Gottesman, S. Micros for microbes: non-coding regulatory RNAs in bacteria. *TRENDS Genet.* **21,** 399–404 (2005).
222. Gripenland, J. *et al.* RNAs: regulators of bacterial virulence. *Nat. Rev. Microbiol.* **8,** 857–866 (2010).
223. Chabelskaya, S., Gaillot, O. & Felden, B. A Staphylococcus aureus small RNA is required for bacterial virulence and regulates the expression of an immune-evasion molecule. *PLoS Pathog.* **6,** 1–11 (2010).
224. Repoila, F. & Darfeuille, F. Small regulatory non-coding RNAs in bacteria: physiology and mechanistic aspects. *Biol. Cell* **101,** 117–131 (2009).
225. Georg, J. & Hess, W. R. cis-antisense RNA, another level of gene regulation in bacteria. *Microbiol. Mol. Biol. Rev.* **75,** 286–300 (2011).
226. Tijsterman, M. & Plasterk, R. H. A. Dicers at RISC: The mechanism of RNAi. *Cell* **117,** 1–3 (2004).
227. Horvath, P. & Barrangou, R. CRISPR/Cas, the immune system of bacteria and archaea. *Science* **327,** 167–170 (2010).
228. Zebec, Z., Manica, A., Zhang, J., White, M. F. & Schleper, C. CRISPR-mediated targeted mRNA degradation in the archaeon Sulfolobus solfataricus. *Nucleic Acids Res.* **42,** 5280–5288 (2014).
229. Hüttenhofer, A., Schattner, P. & Polacek, N. Non-coding RNAs: Hope or hype? *Trends Genet.* **21,** 289–297 (2005).
230. Costa, F. F. Non-coding RNAs: Lost in translation? *Gene* **386,** 1–10 (2007).

231. Mendoza-Vargas, A. *et al.* Genome-wide identification of transcription start sites, promoters and transcription factor binding sites in E. coli. *PLoS One* **4,** (2009).
232. Voskuil, M. I. & Chambliss, G. H. The -16 region of Bacillus subtilis and other gram-positive bacterial promoters. *Nucleic Acids Res.* **26,** 3584–3590 (1998).
233. Hershberg, R. & Petrov, D. A. Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genet.* **6,** (2010).
234. Lamelas, A. *et al.* Serratia symbiotica from the aphid Cinara cedri: A missing link from facultative to obligate insect endosymbiont. *PLoS Genet.* **7,** e1002357 (2011).
235. Yus, E. *et al.* Transcription start site associated RNAs in bacteria. *Mol. Syst. Biol.* **8,** 585 (2012).
236. Lartigue, C. *et al.* Genome transplantation in bacteria: changing one species to another. *Science* **317,** 632–638 (2007).
237. Levin, J. Z. *et al.* Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat. Methods* **7,** 709–715 (2010).
238. Parkhomchuk, D. *et al.* Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res.* **37,** e123 (2009).
239. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping. *Genome Res.* **18,** 1851–1858 (2008).
240. Robinson, J. T. *et al.* Integrative Genome Viewer. *Nat. Biotechnol.* **29,** 24–26 (2011).
241. Maier, T. *et al.* Quantification of mRNA and protein and integration with protein turnover in a bacterium. *Mol. Syst. Biol.* **7,** 511 (2011).
242. Dornenburg, J. E., DeVita, A. M., Palumbo, M. J. & Wade, J. T. Widespread antisense transcription in Escherichia coli. *MBio* **1,** 1–4 (2010).
243. Levine, E., Zhang, Z., Kuhlman, T. & Hwa, T. Quantitative characteristics of gene regulation by small RNA. *PLoS Biol.* **5,** 1998–2010 (2007).
244. Hoops, S. *et al.* COPASI - A COmplex PAthway SImulator. *Bioinformatics* **22,** 3067–3074 (2006).
245. Pich, O. Q., Burgos, R., Planell, R., Querol, E. & Piñol, J. Comparative analysis of antibiotic resistance gene markers in Mycoplasma genitalium: Application to studies of the minimal gene complement. *Microbiology* **152,** 519–527 (2006).
246. Perkins, D. N., Pappin, D. J. C., Creasy, D. M. & Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20,** 3551–3567 (1999).
247. Wodke, J. A. H. *et al.* MyMpn: a database for the systems biology model organism Mycoplasma pneumoniae. *Nucleic Acids Res.* **43,** D618–D623 (2015).
248. Mehta, P., Goyal, S. & Wingreen, N. S. A quantitative comparison of sRNA-based and protein-based gene regulation. *Mol. Syst. Biol.* **4,** 221 (2008).
249. Ortet, P., De Luca, G., Whitworth, D. E. & Barakat, M. P2TF: a comprehensive resource for analysis of prokaryotic transcription factors. *BMC Genomics* **13,** 628 (2012).
250. Pérez-Rueda, E., Collado-Vides, J. & Segovia, L. Phylogenetic distribution of DNA-binding transcription factors in bacteria and archaea. *Comput. Biol. Chem.* **28,** 341–350 (2004).
251. Taniguchi, Y. *et al.* Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science* **329,** 533–538 (2010).
252. Legewie, S., Dienst, D., Wilde, A., Herzel, H. & Axmann, I. M. Small RNAs establish delays and temporal thresholds in gene expression. *Biophys. J.* **95,** 3232–3238 (2008).
253. Shimoni, Y. *et al.* Regulation of gene expression by small non-coding RNAs: a quantitative view. *Mol. Syst. Biol.* **3,** 138 (2007).
254. Levine, E. & Hwa, T. Small RNAs establish gene expression thresholds. *Curr. Opin. Microbiol.* **11,** 574–579 (2008).
255. Aiba, H. Mechanism of RNA silencing by Hfq-binding small RNAs. *Curr. Opin. Microbiol.* **10,** 134–139 (2007).
256. Palazzo, A. F. & Lee, E. S. Non-coding RNA: What is functional and what is junk? *Front. Genet.* **5,** 1–11 (2015).

257. Barré, A., de Daruvar, A. & Blanchard, A. MolliGen, a database dedicated to the comparative genomics of Mollicutes. *Nucleic Acids Res.* **32,** D307–D310 (2004).
258. Djordjevic, S. P. & Tacchi, J. L. in *Mollicutes: Molecular Biology and Pathogenesis* (eds. Browning, G. F. & Citti, C.) (UK: Caister Academic Press, 2014).
259. Tacchi, J. L. *et al.* Post-translational processing targets functionally diverse proteins in Mycoplasma hyopneumoniae. *Open Biol.* **6,** 150210 (2016).
260. Paes, J. A. *et al.* Secretomes of Mycoplasma hyopneumoniae and Mycoplasma flocculare reveal differences associated to pathogenesis. *J. Proteomics* **154,** 69–77 (2017).
261. Miles, R. J., Taylor, R. R. & Varsani, H. Oxygen upake and H2O2 production by fermentative Mycoplasma spp. *J. Med. Microbiol.* **34,** 219–223 (1991).
262. Machado, C. X., Pinto, P. M., Zaha, A. & Ferreira, H. B. A peroxiredoxin from Mycoplasma hyopneumoniae with a possible role in H2O2 detoxification. *Microbiology* **155,** 3411–3419 (2009).
263. Gonchoroski, T. *et al.* Evolution and function of the Mycoplasma hyopneumoniae peroxiredoxin, a 2-Cys-like enzyme with a single Cys residue. *Mol. Genet. Genomics* **292,** 297–305 (2017).
264. Minion, F. C. *et al.* The genome sequence of Mycoplasma hyopneumoniae strain 232, the agent of swine mycoplasmosis. *J. Bacteriol.* **186,** 7123–7133 (2004).
265. Abu-Amero, K. K., Miles, R. J. & Halablab, M. a. Cholesterol protects Acholeplasma laidlawii against oxidative damage caused by hydrogen peroxide. *Vet. Res. Commun.* **29,** 373–380 (2005).
266. Schafer, E. R., Oneal, M. J., Madsen, M. L. & Minion, F. C. Global transcriptional analysis of Mycoplasma hyopneumoniae following exposure to hydrogen peroxide. *Microbiology* **153,** 3785–3790 (2007).
267. Wang, G., Hong, Y., Olczak, A., Maier, S. E. & Maier, R. J. Dual roles of Helicobacter pylori NapA in inducing and combating oxidative stress. *Infect. Immun.* **74,** 6839–6846 (2006).
268. Hwang, F. *et al.* Studies on the phospholipid composition of pathogenic cell membranes of Mycoplasma hyopneumoniae. *FEBS Lett.* **195,** 323–326 (1986).
269. Gardella, R. S. & Del Giudice, R. A. Growth of Mycoplasma hyorhinis cultivar α on Semisynthetic medium. *Appl. Environ. Microbiol.* **61,** 1976–1979 (1995).
270. Lafont, F. & van der Goot, F. G. Bacterial invasion via lipid rafts. *Cell. Microbiol.* **7,** 613–620 (2005).
271. Fürnkranz, U., Siebert-Gulle, K., Rosengarten, R. & Szostak, M. P. Factors influencing the cell adhesion and invasion capacity of Mycoplasma gallisepticum. *Acta Vet. Scand.* **55,** 63 (2013).
272. Chen, J. W. *et al.* Comparative Analysis of Glycoprotein and Glycolipid Composition of Virulent and Avirulent Strain Membranes of Mycoplasma hyopneumoniae. **24,** 189–192 (1992).
273. Romero-García, J., Francisco, C., Biarnés, X. & Planas, A. Structure-function features of a mycoplasma glycolipid synthase derived from structural data integration, molecular simulations, and mutational analysis. *PLoS One* **8,** 1–14 (2013).
274. Holub, B. J. Metabolism and function of myo-inositol and inositol phospholipids. *Annu. Rev. Nutr.* **6,** 563–597 (1986).
275. Castello, A., Hentze, M. W. & Preiss, T. Metabolic Enzymes Enjoying New Partnerships as RNA-Binding Proteins. *Trends Endocrinol. Metab.* **26,** 746–757 (2015).
276. Kim, M. K. & Lun, D. S. Methods for integration of transcriptomic data in genome-scale metabolic models. *Comput. Struct. Biotechnol. J.* **11,** 59–65 (2014).
277. Machado, D. & Herrgård, M. Systematic Evaluation of Methods for Integration of Transcriptomic Data into Constraint-Based Models of Metabolism. *PLoS Comput. Biol.* **10,** (2014).
278. Akesson, M., Forster, J. & Nielsen, J. Integration of gene expression data into genome-scale metabolic models. *Metab. Eng.* **6,** 285–293 (2004).

279. Chowdhury, A., Khodayari, A. & Maranas, C. D. Improving prediction fidelity of cellular metabolism with kinetic descriptions. *Curr. Opin. Biotechnol.* **36,** 57–64 (2015).
280. Karr, J. R. *et al.* A whole-cell computational model predicts phenotype from genotype. *Cell* **150,** 389–401 (2012).
281. Sharma, S., Tivendale, K. A., Markham, P. F. & Browning, G. F. Disruption of the membrane nuclease gene (MBOVPG45_0215) of Mycoplasma bovis greatly reduces cellular nuclease activity. *J. Bacteriol.* **197,** 1549–1558 (2015).
282. Minion, F. C., Jarvill-Taylor, K. J., Billings, D. E. & Tigges, E. Membrane-associated nuclease activities in mycoplasmas. *J. Bacteriol.* **175,** 7842–7847 (1993).
283. Siqueira, F. M. *et al.* New insights on the biology of swine respiratory tract mycoplasmas from a comparative genome analysis. *BMC Genomics* **14,** 175 (2013).
284. Yamamoto, T., Kida, Y., Sakamoto, Y. & Kuwano, K. Mpn491, a secreted nuclease of Mycoplasma pneumoniae, plays a critical role in evading killing by neutrophil extracellular traps. *Cell. Microbiol.* **19,** 1–11 (2017).
285. Cacciotto, C. *et al.* Mycoplasma agalactiae MAG_5040 is a Mg2+-Dependent, Sugar-Nonspecific SNase Recognised by the Host Humoral Response during Natural Infection. *PLoS One* **8,** 1–11 (2013).
286. Ishag, H. Z. A. *et al.* A replicating plasmid-based vector for GFP expression in Mycoplasma hyopneumoniae. *Genet. Mol. Res.* **15,** 1–8 (2016).
287. Ceriani, R. *et al.* Group contribution model for predicting viscosity of fatty compounds. *J. Chem. Eng. Data* **52,** 965–972 (2007).
288. Cevc, G. How Membrane Chain-Melting Phase-Transition Temperature Is Affected by the Lipid Chain Asymmetry and Degree of Unsaturation: An Effective Chain-Length Model? *Biochemistry* **30,** 7186–7193 (1991).
289. MycoSynVac Engineering Mycoplasma pneumoniae as a broad-spectrum animal vaccine. Available at: http://www.mycosynvac.eu/.

Summaries and acknowledgements

Summary

Mycoplasma hyopneumoniae (*M. hyopneumoniae*) is a bacterial pathogen that has evolved from a gram-positive ancestor and specifically colonizes the lower respiratory tract of pigs where it causes enzootic pneumonia and plays a major role in the development of respiratory disease in pigs. Whole-cell inactivated vaccines are available that lower the severity of disease and are widely applied in pig industry to prevent clinical signs and improve pig herd health. However, production of these vaccines is challenging because it is not known which bacterial components are needed for protection and complex cultivation media are needed because growth requirements are not completely understood. The aim of this thesis was to understand growth and survival strategies of *M. hyopneumoniae* during infection, to integrate this knowledge with metabolic modeling under conditions used for vaccine production and apply this knowledge to improve the current production process for *M. hyopneumoniae* vaccines.

Chapter 1 provides a general introduction into the disease, treatment and prevention methods with a focus on vaccines. I then introduce the characteristics of the *M. hyopneumoniae* genome, transcriptome and review the current knowledge on infectious mechanisms and the response of the pig to infection and vaccination. Finally, I discuss the challenges related to vaccine production and introduce systems biology tools that will be applied in the thesis. In **chapter 2** we define a strategy for risk-based process development of bacterial vaccines which provided the framework for future studies performed during this thesis project. We propose to integrate the academic workflow for rational strain design with the industry standard for process design. Systems biology tools, especially genome-scale metabolic models, play an essential role in this strategy because application of these tools reduces process risks and increases process understanding. Therefore, in line with this strategy, we created a manually curated genome-scale metabolic model of *M. hyopneumoniae* which we applied to dynamically model the cultivation step in the vaccine production process (**chapter 3**). We found that only 16% of cellular energy in a standard fermentation was used for growth and 84% was used for non-growth associated maintenance. By model-driven experimentation we were able to increase the fraction of cellular energy used for growth by addition of pyruvate to the production medium, and showed in dedicated fermentor experiments that the improved process reached a 2.3 times higher biomass yield. Although the metabolic model helped to increase process yield, it did not allow prediction of a defined cultivation medium without components from porcine origin. Therefore, to better understand the dependency of *M. hyopneumoniae* on host derived components, we performed a functional comparison of 80 mycoplasma genomes and used multivariate and machine-learning algorithms to relate functional capability to the specific host and niche of mycoplasma species (**chapter 4**). This analysis allowed us to identify protein domains possibly needed for growth and survival in the pig lung. In addition, we found that protein domains expected to be essential for bacterial growth were not persistently present in mycoplasma genomes suggesting that alternative domain

configurations exist that bypass their essentiality. To better understand whether the proteins we identified as possibly important for survival in pigs actually play a role during *M. hyopneumoniae* infection, we sequenced the bacterial mRNA during infection in **chapter 5** and compared the *in vivo* transcriptome to that of broth grown mycoplasma. We found up-regulation during infection of F_1-like ATPase and several genes related to nucleotide metabolism, spermidine transport and glycerol-3-phoshate transport. Genes related to cilium adhesion, glycerol uptake, cell division and myo-inositol metabolism were found down-regulated *in vivo*. In our analysis we also paid specific attention to the role of non-coding RNAs (ncRNAs) as these were found to play an important role during infection in other bacterial pathogens and we identified differentially expressed ncRNAs during infection. In **chapter 6** we build upon our mycoplasma basis to further analyse the role of ncRNAs in bacterial genomes. We identified an exponential relationship between the AT content of genomes and the number of ncRNAs and propose that this relation is the result of spurious transcription, which is more likely to occur in AT rich genomes. This hypothesis is further substantiated by showing that spurious transcription demands minimal cellular energy and that overexpression of cis-binding ncRNAs in *M. pneumoniae* did not influence the level of proteins translated from their overlapping mRNAs. Finally, in **chapter 7** I discuss four system strategies, identified in this thesis and derived from recent literature, and discuss how these strategies could be integrated in the metabolic model of *M. hyopneumoniae*. Lastly, I provide an outlook on the next steps needed for improvement of the production process for *M. hyopneumoniae* vaccines.

In conclusion, this work provided novel insight in the metabolic capability of *M. hyopneumoniae* based on the proteome domain content, captured in a genome-scale metabolic model and studied under *in vitro* and *in vivo* conditions. Biomass yield of the cultivation step for vaccine production was increased and the basis was laid to further improve the production process for *M. hyopneumoniae* vaccines using model-based experimentation.

Samenvatting

Longontsteking is een veel voorkomende ziekte in varkens, die kan worden veroorzaakt door verschillende pathogenen. Vatbaarheid voor infectie wordt bepaald door de leeftijd van het varken, genetische achtergrond en verschillende omgevingsfactoren. Als biggen meer dan tien weken oud zijn, wordt longonsteking voornamelijk veroorzaakt door: "Porcine Reproductive and Respiratory Syndrome Virus (PRRSV)", "Porcine Circo Virus (PCV)" en de bacterie *Mycoplasma hyopneumoniae* (*M. hyopneumoniae*). *M. hyopneumoniae* is een obligaat pathogeen en veroorzaakt enzoötische pneumonie, een aandoening die in de stal herkenbaar is aan een droge, persistente hoest. Besmettingen met *M. hyopneumoniae* komen veel voor en de bacterie verspreidt zich makkelijk via neus-tot-neus contact en aërosolen. Vaccinatie of behandeling met antibiotica, verminderen de ziekteverschijnselen, maar dit leidt niet tot eradicatie van de ziekteverwekker.

Mycoplasma infecteren specifiek een gewervelde gastheer, en tijdens co-evolutie met de gastheer, is het bacteriële genoom aanzienlijk gereduceerd. Dit heeft ertoe geleid dat mycoplasma soorten het kleinste bekende genoom bevatten, van organismes die zich kunnen repliceren zonder gebruik te maken van een gastheercel. Mycoplasma bacteriën zijn te kweken op complexe media met toevoegingen van dierlijke peptonen en serum. *In vivo*, bindt *M. hyopneumoniae* zich specifiek aan de trilharen (cilia) aanwezig op de epitheelcellen van de luchtwegen (luchtpijp, bronchi, bronchioli). Aanwezigheid van *M. hyopneumoniae* veroorzaakt een ontstekingsreactie die zichtbaar wordt door de vorming van laesies. Histologisch is in preparaten van longweefsels influx van neutrofielen, lymfocyten en macrofagen zichtbaar. Schade aan longweefsel wordt waarschijnlijk veroorzaakt door bacteriële virulentiefactoren en de reactie daarop van het immuunsysteem van het varken. De kennis van het infectieproces is echter niet volledig. Zo is het nog niet bekend welke virulentiefactoren essentieel zijn voor het veroorzaken van infectie.

Omdat in dit geval niet bekend is welke virulentiefactoren ziekte veroorzaken, worden vaccins gebruikt die volledige geïnactiveerde bacteriën bevatten. De productie van deze vaccins wordt bemoeilijkt doordat: i) de component die bescherming biedt niet bekend is en daarom opbrengst en de kwaliteit van het entmateriaal moeilijk te bepalen is, ii) productietechnische variatie hoog is door het gebruik van complexe dierlijke componenten in kweekmedia en iii) dierlijke componenten duur zijn en mogelijk ziekteverwekkers bevatten die geïnactiveerd moeten worden voordat ze als grondstof gebruikt kunnen worden. Dit proefschrift had als primair doel om de groeikarakteristieken en overlevingsmechanismen van *M. hyopneumoniae* tijdens infectie en tijdens groei in het kweekmedium beter te begrijpen en te modeleren. Vervolgens is deze kennis gebruikt om in een model gedreven aanpak het productieproces voor vaccins te verbeteren.

In **hoofdstuk 1** geef ik een inleiding op het proefschrift. Vervolgens wordt in **hoofdstuk 2** de problematiek geanalyseerd die speelt bij de productie en ontwikkeling van bacteriële vaccins. Vaak zijn in deze vaccins de beschermende componenten niet bekend en is er variatie in productieprocessen door het gebruik van grondstoffen van dierlijke oorsprong. Wij stellen voor om in het vaccin ontwikkelingstraject op meerdere momenten product risicoanalyse toe te passen en daarbij gebruik te maken van de kennis en inzichten verkregen met metabole modellen van bacteriën, aangevuld met analyse van de bacterie met systeem brede omics technieken, om de kans op het bereiken van een procesverbetering te vergroten. De methodiek die wij voorstellen past in de huidige standaard in de industrie en past bij de stappen die gebruikelijk zijn in academische

onderzoekstrajecten. In **hoofdstuk 3** zetten we de eerste stap in deze aanpak, door een metabool model te maken van *M. hyopneumoniae*, en het model te gebruiken om beter inzicht te krijgen in de energiebalansen in de bacterie. We tonen aan dat tijdens groei in cultures 84% van de cellulaire energie naar onderhoudstaken gaat in de cel die niet gerelateerd zijn aan groei. Om deze balans te verbeteren en daarmee meer groei te bewerkstelligen, hebben we vervolgens met modelsimulaties onderzocht welke toevoegingen aan medium extra energie kunnen leveren voor groei. We laten zien dat toevoeging van pyruvaat aan het medium een verbetering geeft van de groeisnelheid en resulteert in een 2.3x hogere opbrengst van biomassa in cultures.

De studies met het model stelden ons niet in staat om de samenstelling van het medium verder te vereenvoudigen. Om meer inzicht te krijgen in de functionele eigenschappen van *M. hyopneumoniae* in relatie met de gastheer, hebben we in **hoofdstuk 4** een genoom vergelijking gedaan tussen 80 mycoplasma soorten uit verschillende fylogenetische groepen, om bacteriële functies te identificeren die belangrijk zijn voor kolonisatie van een gastheer of van een specifieke niche in een gastheer. Hiervoor hebben we de 80 mycoplasma genomen opnieuw geannoteerd op basis van eiwit domeinen en laten zien dat de 80 beschibare genomen voldoende informatie bevatten voor een sluitende analyse omdat het totaal aantal unieke eiwitdomeinen, het "pan-domeinome", een constante waarde bereikt. Door gebruik te maken van algoritmes die mycoplasma species kunnen toewijzen in groepen aan de hand van de gastheer of de niche, konden we op basis van de domein samenstelling voorspellen of een mycoplasma een varken, rund of mens als gastheer heeft en welke domeinen van belang zijn voor het toewijzen van die gastheer. In onze analyse hebben we ook het minimale genoom meegenomen, JCVI-Syn3.0, dat recent gemaakt is op basis van het genoom van *Mycoplasma mycoides*. In deze analyse viel op dat functies waarvan we verwachten dat ze essentieel zijn omdat ze voorkomen in het minimale genoom, niet altijd voorkomen in andere mycoplasma en dus vervangbaar zijn door alternatieve functies.

Om te onderzoeken welke genen van belang zijn voor groei of overleving van *M. hyopneumoniae* tijdens infectie in het varken, hebben we in **hoofdstuk 5** het *in vivo* transcriptoom van de bacterie bepaald tijdens groei in de long, met behulp van "RNA sequencing". Voor deze analyse moet de concentratie van bacterieel mRNA in verhouding tot varkens mRNA voldoende hoog zijn om betrouwbare data van de bacterie te krijgen. We laten zien dat het opwerken van RNA uit geïnfecteerd longweefsel niet voldoende bacterieel RNA oplevert en hebben vervolgens een methode ontwikkeld waarbij met een spoeling van een geïnfecteerd gedeelte van de long, en vervolgens een verwijdering van mRNA van het varken, een voldoende hoge verhouding tussen bacterieel en varkens mRNA wordt verkregen. Het vergelijk van het bacteriële transcriptoom tijdens groei in het varken met het transcriptoom tijdens groei in kweekmedium liet zien dat in totaal 62 genen differentieel tot expressie kwamen tijdens infectie. Genen die hoger tot expressie kwamen tijdens infectie coderen voor nucleases, glycerol-3-phosphaat transport en spermidine transport. Expressie van genen die coderen voor eiwitten gerelateerd aan glycerol transport, adhesie aan de cilia en algemene functies nodig voor celdeling was lager tijdens infectie. In deze studie hebben we ook geanalyseerd of er differentiële expressie was van kleine, niet voor eiwitten coderende RNAs (ncRNAs) en vonden 28 omhoog gereguleerde en 66 geremde ncRNAs. De functie van deze ncRNAs in *M. hyopneumoniae* is nog niet bekend. Om meer inzicht te krijgen in de rol van ncRNAs in mycoplasma species hebben we in **hoofdstuk 6** een analyse gedaan van het aantal ncRNAs in verschillende mycoplasmas en dat vergeleken met andere bacteriën en een chloroplast. Dit onderzoek liet zien dat er een

exponentieel verband was tussen de hoeveelheid ncRNAs en het percentage AT-baseparen in het genoom. Onze verwachting was dat veel van deze ncRNAs gemaakt worden door een ongecontroleerde start van transcriptie als gevolg van de hoge concentratie AT in het genoom. Om deze hypothese te onderbouwen laten we zien dat de totale energie die nodig is voor vorming van de ncRNAs laag is en tevens dat overexpressie van ncRNAs in *M. pneumoniae* geen invloed heeft op de concentraties van de eiwitten afgelezen van de mRNAs waaraan de ncRNAs zouden kunnen binden.

Als laatste bediscussieer ik in **hoofstuk 7,** een aantal vindingen in deze thesis en plaats ze in het licht van recent onderzoek en dat doe ik aan de hand van een viertal groeistrategieën die voor *M. hyopneumoniae* van belang kunnen zijn, maar nog niet meegenomen waren in het metabole model. Vervolgens bediscussieer ik hoe deze kennis geïntegreerd kan worden in het metabole model en wat de vervolg stappen zijn om het productie proces voor *M. hyopneumoniae* vaccins verder te verbeteren.

Samenvattend concluderen we dat tijdens groei van *M. hyopneumoniae* in cultures slechts 16% van de cellulaire energie gebruikt wordt voor groei maar dat dit percentage verbeterd kan worden door toevoeging van pyruvaat aan kweekmedia waardoor een hogere groeisnelheid en hogere concentratie biomassa wordt verkregen. Met behulp van RNA sequencing hebben we aangetoond dat er tijdens infectie genen gerelateerd aan glycerol-3-fosfaat transport, nuclease activiteit en spermidine transport tot overexpressie komen en daarom mogelijk een rol spelen tijdens infectie. Daarnaast tonen we aan dat ncRNAs in mycoplasma species grotendeels het resultaat zijn van ongecontroleerde transcriptie. Onze analyse van eiwitdomeinen laat zien dat het pan-domeinome van het mycoplasma genus gesloten is en dat er alternatieve domeinsamenstellingen mogelijk zijn voor essentiële functies in minimale genomen. Het metabole model beschreven in dit proefschrift biedt de basis voor verder onderzoek naar metabole netwerken in *M. hyopneumoniae* en zal een belangrijke rol spelen bij verdere verbetering van het vaccin productieproces.

Dankwoord / Acknowledgements

Het PhD avontuur zit erop! Na bijna vijf jaar ben ik toegekomen aan het laatste hoofdstuk van dit boekje. Ik ga een poging doen een ieder te bedanken die mij de afgelopen jaren heeft geholpen met dit promotieonderzoek; door mij direct te ondersteunen in de wetenschappelijke werkzaamheden; door mij te helpen met andere projecten bij MSD of door mij te ondersteunen buiten het werk om.

Peter, mijn co-promotor. Het was een beetje zoeken in het begin naar een constructieve wijze van samenwerken. Ik zag soms de toegevoegde waarde niet altijd van een deelonderzoek maar uiteindelijk valt alles mooi samen in dit proefschrift. Bedankt voor alle reviews en discussies, ik vind het knap dat je terwijl je een grote groep AIO's begeleidt toch altijd weer een nieuwe invalshoek weet te vinden voor de verschillende projecten en die dan ook duidelijk kan uitleggen.

Jetta, tevens mijn co-promotor. Bedankt voor alle tijd en energie die je in mijn begeleiding hebt gestoken. Zonder jouw hulp had ik dit boekje niet op tijd, en niet op deze wijze af kunnen krijgen. Jij weet als geen ander wat er nodig is om een wetenschappelijke publicatie goed neer te zetten en tevens wat er nodig is om een promotieproject tot een succesvol einde te brengen. Ik heb veel van je geleerd de afgelopen jaren en hoop dat te blijven doen tijdens gezamelijke projecten bij MSD.

Simen-Jan, de eerste dag dat je manager werd van BTS-B werd je geconfronteerd met een reorganisatie en ging ik het merendeel van mijn tijd aan mijn promotieproject werken. Dat was een uitdaging en gedurende de jaren daarna was het altijd zoeken naar een goede balans tussen promotiewerk en andere projecten voor MSD. Bedankt dat je mij altijd bent blijven steunen en steeds nauw betrokken bent geweest bij het project. Je had ook altijd oog voor de menselijke kant van het werk en daardoor weet ik nu dat als ik niet meer sport door werkdruk er iets goed mis is!

Paul, bedankt voor je steun gedurende het opstarten van dit project en het scheppen van de randvoorwaarden waardoor ik aan mijn promotieonderzoek kon werken in een internationale omgeving met vele specialisten. Ik heb je inzicht gedurende onze werkoverleggen en brainstorms altijd erg gewaardeerd.

Vitor, mijn promotor. Bedankt voor je steun gedurende onze routine overleggen en het leggen van de nodige contacten binnen de wetenschappelijke gemeenschap die dit onderzoek verder hebben gebracht.

Ik wil de leden van de leescommissie, Prof. Huub Savelkoul, Prof. Bas Teusink, Prof. Michel Eppink en Dr. Florence Tardy graag bedanken voor het kritisch lezen en beoordelen van dit proefschrift.

Er zijn een aantal mensen binnen MSD Animal Health die mij vanaf het begin hebben gesteund en daardoor dit promotieonderzoek mogelijk hebben gemaakt. **Edwin**, bedankt dat ik deze kans heb gekregen. Jij bent in staat om het beste in mensen naar boven te halen! **Jos**, vanaf het eerste moment dat het plan om te promoveren naar voren kwam heb jij mij

altijd gesteund, een grotere investering in mijn ontwikkeling had ik mij niet kunnen wensen, bedankt voor je vertrouwen. **Sjo**, bedankt voor je inspanningen tijdens het opstarten van dit onderzoek en de steun in de jaren daarna. **Frank**, bedankt dat je als interim afdelingshoofd dit project hebt gesteund en voor alle steun en motivatie in de jaren daarna. Ik heb de PhD Comics nog steeds liggen die je mij bij de start hebt gegeven, inmiddels is het bijna nostalgie.

Zonder de steun van ons sterke BTS-B team in Boxmeer had ik dit promotieonderzoek niet kunnen doen. **Marian**, wat hebben wij een hoop mycoplasma projecten gedraaid de afgelopen jaren! Ik waardeer je inzet enorm en je was ook altijd bereid om werkzaamheden van mij over te nemen, zelfs als die klassiek bij een projectleider zouden moeten liggen. **Maud**, als D&Ter binnen ons team heb je je waarde voor BTS al snel bewezen. Bedankt voor je inspanningen voor de qPCR, die resultaten gaan we zeker nog publiceren! **Marloes**, "miss mycoplasma" (?), ik hoop dat je het stokje een keer van mij over wilt nemen. **Henriëtte**, bedankt voor het praktisch begeleiden van de stagiaires op dit project en je steun als het een keer tegenzat op het lab. **Jenneke, Will, Remco** en **Marieke**, bedankt voor jullie steun de afgelopen jaren. **Harald, Jeoffrey, Richard, Frank, Simen-Jan,** bedankt voor jullie steun en bedankt dat jullie mij uit de wind hebben gehouden de afgelopen jaren zodat ik mij op het onderzoek kon richten.

Gedurende mijn promotie heb ik ook veel mogen samenwerken en steun gehad van andere afdelingen binnen MSD Animal Health zoals **Microbiologische R&D, ASD, D&T** en **BTS-V**. Ik ga niet iedereen hier noemen maar ik waardeer de interesse, collegialiteit en de bereidheid om te helpen enorm! Specifiek wil ik graag **Ruud** bedanken voor het kritisch lezen van mijn stukken en een niet aflatende interesse in mijn werkzaamheden, vanaf het eerste moment dat ik bij MSD binnenkwam. **Maarten**, bedankt voor al je steun de afgelopen jaren, ik hoop toch voor jouw (of mijn) pensioen een ACF *M. hyo* medium op te leveren. Maurice, bedankt voor al je hulp tijdens de verschillende *M. hyo* studies. **Bartjan**, bedankt dat we met onze (q)PCR vragen altijd bij je terecht konden. **Pieter**, we vormden een geweldig duo voor de longspoelingen, bedankt dat je altijd bereid was extra tijd in te plannen zodat ik mijn lastige samples kon verzamelen. **Mieke, Patricia**, bedankt voor jullie hulp met het genereren en de interpretatie van de histologie samples. **Eveline**, bedankt voor je steun, hulp met de planning en het laten zien dat er licht is aan het einde van de tunnel.

Voor alle administratieve steun en organisatorische hulp ben ik veel dank verschuldigd aan **Lieke, Carien** en **Annemarie** in Boxmeer en **Carolien** en **Mirella** in Wageningen.

John, bedankt voor het tijdig en kritisch reviewen van mijn stukken en voor de interessante discussies en je scherpe vragen n.a.v. het onderzoek.

De studenten die mij geholpen hebben met dit onderzoek wil ik graag van harte bedanken. **Niels**, jij kwam met een flinke dosis ervaring vanuit de Bacto bij ons binnen en was daardoor in staat zelfstandig aan je onderzoek te werken. De Tier 9's zijn binnen het bedrijf nooit meer overtroffen en ik ben blij dat je na je stage weer een plekje hebt gevonden binnen BTS. **Luuk**, jij was op papier als student scheikundige technologie toch een beetje vreemde eend tussen de biotechnologen, vanaf de eerste dag heb je echter laten zien dat je je prima thuisvoelde en kon functioneren in deze omgeving. Jouw onderzoek heeft laten zien hoe complex een complex medium kan zijn en hoe moeilijk het vinden van een

enzymatische activiteit kan zijn, dit vormt nog altijd een basis voor het vervolgonderzoek. **Sacha**, ik noem vaak metabole capaciteit in dit proefschrift maar jouw experimenten hebben laten zien wat dat echt inhoudt. Bedankt voor je scherpe blik en doorzettingsvermogen.

I would like to thank the PhD's and postdocs in Wageningen that made my time there really enjoyable. Every Tuesday I entered into a different dimension with shifted social standards and conversation topics beyond everyone's imagination. **Mark**, thanks for being first a student on the project already in 2009 and then a great support during the first two years of my PhD. I used your AWK scripts until the end of the project and still only moderately understand how they work. Jasper, thanks for helping me with an infinite amount of SPARQL queries. I would like to thank all my roommates: **Rienk, Nikolaos, Emma, Rob, Benoit and Maarten**. **Maarten**, thanks for being just the most magnificent human being that ever walked this planet. **Benoit**, thanks for teaching me the art of procrastination and joining me for a shear infinite amount of Chinese take-away lunches. **Bastian**, we started on the same day in SSB; I think you showed over the past years that the title "PhD student of the year" was well deserved. **Niels**, thanks for all the discussions on mycoplasma metabolomics and being my MycoSynVac buddy from the start of the project. I'm thankful for the support from the inhabitants of the most socially gifted office in SSB: **Niru, Nhung, Linde and Ruben**. **Linde** if you maintain your positive energy, your PhD and future career will be a blast. **Nhung**, thanks for showing me algae can be food; at least this means they're somehow useful. **Bart**, thanks for all the IT support.

Maria, thanks for your help with and input for chapter 3 and in general your help with data interpretation.

I would like to thank the researchers in the MycoSynVac project for the great support and interesting collaborations over the years. I highly value the meetings that we had and especially like to thank **Maria, Pascal** and **Luis** for their support.

Jos, bedankt voor het geweldige grafisch ontwerp, soms zeggen plaatjes echt meer dan duizend woorden.

Naast werk was er gelukkig toch nog wel tijd om Bottendaal onveilig te maken. **Albert** en **Chris**, bedankt voor de vele potjes pool, rondjes fietsen en herstelbiertjes na afloop. Poolpikkies for life! **Annique** en **Dave**, bedankt voor de geweldige avonden op het trappetje en de vele spelletjesavonden. Natuurlijk ook veel dank aan mijn **BvdB huisgenootjes**, dit boekje is hét bewijs dat ik vaccins maak, voor dieren...

Sneeuwhappen op de skipiste vormde de jaarlijkse harde reset aan het begin van het jaar. **Nancy, Hans, Jorg, Jorien, Albert, Marlies**, bedankt voor de epische vakanties en ik hoop dat er nog veel zullen volgen.

Melanie, bedankt voor je steun en de vele gezellige etentjes en spelletjesavonden. **Jasper**, volgens mij kan jij mijn eerste paper beter uitleggen dan ik dat kan, bedankt voor je steun en de therapeutische hardloopsessies. **Rolf**, onder het genot van een biertje moeten we nog maar eens rustig door dit boekje heenlopen, ik ben erg benieuwd naar jouw mening over de onderzoeken, bedankt voor vele jaren vriendschap en steun.

Mijn paranimfen, ik zou niet weten wat ik zonder ze zou moeten doen. **Marlies**, oneindig veel opbeurende telefoontjes, weekendjes en vakanties samen. Je hebt vanaf het begin in mij geloofd en bent een onmisbare steun geweest. Ik ben trots dat jij als paranimf naast me staat! **Albert**, toen ik tien jaar geleden in Nijmegen kwam wonen had ik me niet kunnen voorstellen dat ik mijn PhD gevoelsmatig zou afsluiten met een metal meeting in Eindhoven. Hieraan zijn natuurlijk een stroom van concertjes, festivals, caravanparty's en avondjes in de kroeg voorafgegaan. Bedankt dat je mij hierin hebt meegenomen en voor de vriendschap en steun de afgelopen jaren, ik ben blij dat je tijdens de verdediging ook weer naast mij staat.

Mijn familie in Brazilië wil ik graag bedanken voor de steun: **tante Anneke, oom Juan, Carlitos, Alejandro** en **Marielba**. Ik hoop nu dit werk is afgerond snel een keer jullie kant op te komen. **Tante Anne**, bedankt voor de steun de afgelopen jaren.

Er is niets op deze wereld belangrijker voor mij dan mijn broers: **Remko, Fenno, Guido** en **Sergio**, bedankt voor jullie steun en ik hoop de komende tijd weer wat socialer te kunnen doen en vaker op bezoek te komen. Ook veel dank aan jullie lieve vriendinnetjes: **Iris, Joske** en **Linda**, voor hun steun.

Pa en **ma**, jullie hebben mij altijd gesteund om het beste uit mijzelf te halen maar mij wel altijd de vrijheid gegeven om zelf te bepalen hoe ver ik daarin wilde gaan. Aan de basis van dit onderzoekswerk ligt nieuwsgierigheid, door willen vragen, en dat heb ik van huis uit van jullie meegekregen. Bedankt voor alle steun de afgelopen jaren in de momenten dat het moeilijk was. Zonder jullie steun was dit boekje er niet geweest.

Tjerko Kamminga

September 2017

About the author

List of publications
Training activities
Curriculum Vitae

List of publications

Tjerko Kamminga, Vitor Martins dos Santos, Jetta J.E. Bijlsma and Peter J. Schaap Transcriptome sequencing shows up-regulation of F_1-like ATPase and down-regulation of the P102 cilium adhesin in *Mycoplasma hyopneumoniae* during infection. *Manuscript prepared for submission*

Tjerko Kamminga, Simen-Jan Slagman, Vitor A.P. Martins dos Santos, Jetta J.E. Bijlsma and Peter J. Schaap Risk-based bioengineering strategies for reliable bacterial vaccine production. *Submitted for publication*

Tjerko Kamminga, Simen-Jan Slagman, Jetta J.E. Bijlsma, Vitor A.P. Martins dos Santos, Maria Suarez-Diez and Peter J. Schaap Metabolic modeling of energy balances in Mycoplasma hyopneumoniae shows that pyruvate addition increases growth rate. *Biotechnol. Bioeng.*, 2339–2347 (2017).

Tjerko Kamminga, Simen-Jan Slagman, Jetta J.E. Bijlsma, Vitor A.P. Martins dos Santos, Maria Suarez-Diez and Peter J. Schaap Persistence of Functional Protein Domains in Mycoplasma Species and their Role in Host Specificity and Synthetic Minimal Life. *Front. Cell. Infect. Microbiol.* **7**, 31 (2017).

Verónica Lloréns-Rico, Jaime Cano, **Tjerko Kamminga**, Rosario Gil, Amparo Latorre, Wei-Hua Chen, Peer Bork, John I. Glass, Luis Serrano and Maria Lluch-Senar Bacterial antisense RNAs are mainly the product of transcriptional noise. *Sci. Adv.* **2**, 1–10 (2016).

Training activities

Discipline specific activities	**Year**
MBI module: Downstream Processing	2009
Basic modelling Biologists	2014
Lorentz Workshop - Integrative Cell Models	2015
MSD Science&Technology days	2016
IOM2016 conference	2016
Bacterial and Vaccine manufacturing meeting	2015
General courses	
Presentation Skills	2016
Effective behaviour	2013
Effective communication	2008
Merck Sigma Green Belt	2011
Optionals	
PhD meetings SSB & Microbiology WUR (MIB/SSB)	2013-2017
Bioinformation Technology	2014
Molecular Systems Biology	2013
Writing of proposal	2013

Curriculum Vitae

Tjerko Kamminga was born on the 30th of August 1983 in 's-Gravenhage (The Hague), The Netherlands. Fascinated by technology as a child but only moderately skilled in physics and mathematics he decided to opt for a career in Biotechnology, a field that was predicted to become "booming" and hence should offer great career opportunities. In 2001, after finishing his high-school education in Zoetermeer, he started his studies at Delft University of Technology. Mainly interested in the process-side of biotechnology he fulfilled internships in the Industrial Microbiology group at Delft University and at Xendo Manufacturing B.V. He did his graduation project in the Bioprocess Technology group at Delft University which was aimed at understanding short-term dynamics of glycolysis in *Saccharomyces cerevisiae*. Highly motivated to start a career in pharmaceutical industry he applied for a position at Intervet International B.V. (now MSD Animal Health) in Boxmeer and was hired as scientific assistant in October 2007 to work on the development and scale-up of production processes for bacterial vaccines. After needing almost six years to master the Brabant and Limburg accents, he was offered the opportunity to pursue a PhD degree by performing studies aimed at acquiring better understanding of *Mycoplasma hyopneumoniae* metabolism and physiology. This research project started in January 2013 and was performed in collaboration with the Systems and Synthetic Biology group at Wageningen University and the results are described in this thesis.
In July 2017, Tjerko continued his career at MSD Animal Health in Boxmeer as senior scientist bioprocess technology and support and works on the development of production processes for animal vaccines.

The majority of this work was funded by MSD Animal Health, Bioprocess Technology & Support, Boxmeer, The Netherlands.

In addition, financial support was received from the European Union's Horizon 2020 research and innovation program under grant agreement No. 634942.

Cover design: Jos van Heeswijk, www.studiojos.com
Printed by: Proefschriftmaken.nl || The Netherlands