Engineering Mycoplasma species for biotechnological and biomedical applications

Ariadna Montero-Blay

TESI DOCTORAL UPF / 2021

Thesis supervisors
Prof. Luis Serrano Pubul & Dr. Carlos Piñero-Lambea
EMBL/CRG Systems Biology Research Unit

Centre for Genomic Regulation (CRG)

upf. Universitat Pompeu Fabra Barcelona

To all the women that for decades have been silenced and under-evaluated. Your fight, bravery, and determination deserve my eternal gratitude.

Dedication and Acknowledgments

This thesis has been an in-depth learning experience for me personally and professionally.

I want to acknowledge Luis Serrano, my supervisor, my mentor. From the very first day until the very last, you have been by my side. You have a brilliant mind, and you have a special touch for recruiting friendly and smart people for your team. Even without any scholarship, you allow me to join the lab to push for my ideas. You have boosted my creativity and ambition, and always, even in your busy days, you have had time for anything I needed. It has been a pleasure working close to you, I learned a lot from your wise mind in a humble heart. I will always be grateful to you, and I will conserve everything you teach me for the rest of my life.

Carlos, you have taught me everything I should need to work in a laboratory from scratch. I remember the day you spent more than 3 hours explaining to me how to design a Gibson cloning, and now I have more than 300 in my collection. You have curiosity, patience and you are a great docent. You have been by my side in the first part of my thesis. You help me learn how to write scientifically and present my work orally, which you know that I was not comfortable with initially. I sincerely appreciate you a lot. Thank you for listening to me and always respect my criteria. I was fortunate the day you decided to supervise someone.

I would like to extend my gratitude to my thesis committee members, Manuel Irimia, Fátima Gebauer and Marc Güell. You always have been supportive, you have guided me in the different projects, and more importantly, you have given precious career and science-path management advice. Thank you for everything.

Maria Lluch, you are a clear example for me about perseverance, ambition. You have taught me how dreams can be true. Congratulations on everything you are achieving. You deserve it thoroughly. Samuel Miravet, you are adorable, always helpful, and you scientifically improve everything you touch. You have a clever mind, and you will find the position you deserve. Claire Lastrucci, it was an authentic pleasure to work with you. You give me precious advice, and you have a special light. Thank you for everything. I wish you the best with the new career path you are building. Sarah di Bartolo, we had not worked directly together, but you have been a reference during the whole Ph.D. I always ask for your opinion, advice. You are frank, brilliant, and a marvelous mentor. I am fortunate to have you in the lab while I was doing my thesis. Irene Rodriguez, you came with fresh air to the laboratory and I genuinely believe we can do unique projects together. Javier Delgado, you are a brilliant chemist. You have been supportive, caring, and always with an excellent attitude to collaborate. This has only been the beginning of something ambitious but unique. You're honest, and we need more people in life as brave and honest as you are. Thank you for everything. Damiano and Leandro, you both are great scientists and even more great people. You have always been open to collaborate, open to discuss, and very helpful when I needed it. Hannah, thank you for always being so sensitive; your pieces of advice are gold, you have a very kind heart in a beautiful and clever mind. Ludovica, we have been through this thesis together. We had suffered and enjoyed. We have help each other to overcome all situations. You are a true friend of mine. Xavi, your attitude as labmate is a tresure,

and scientifically, you shine. Raul Burgos, you know everything about Mycoplasma, and I admire from you the fact of always going beyond a scientific question.
Miquel, Alicia, Carolina Segura, Daniel Shaw, Eva Garcia, Daniel Gerngross, Rocco, Sira, Violeta, Tony, Eva Yus, Carolina Gallo, Martin, Jae-Seong, Laia you have contributed a lot to my career development. You are a great group of people. Reyes, Magalí, you are very kind and always resolutive, thank you.
For all the CRG Community, it has been a pleasure to meet at some point and discuss career and life with all of you.

Irene, Isa, and Francesca. We met on the first day of our PhDs, but our friendship will remain for the rest of our lives. You have been my family in Barcelona. I love you

Laia, Clara, Elisa, Marcos, Marc, Jorge. Nos conocimos en el master y años más tarde, aquí seguimos todos unidos. Edu, Elena, Kati y todos los excelentes científicos y personas que me he ido encontrando.

Sandra, Yaiza, Maria, Nuria and the rest of BCN Ladies, gracias por hacerme olvidarme de la ciencia, por ser tan estupendas y por pasarlo tan bien. Os admiro a cada una de vosotras.

Júlia, Laura, Barbi, Nina and Game of Troncas. Haveu aconseguit que els dimarts siguen els millors dies de la setmana. Júlia, ets un referent personal i professional. Us admiro molt.

A les meues amigues de tota la vida, a les Poblanes, a les guapes i les listes. Sou el millor regal que mai m'ha fet ningú, perquè passen els anys i continuem juntes, en les bones i males. Perquè no existeixen els 360 km que ens separen i perquè cadascuna, cada dia, em demostra el vertader valor de la paraula amistat.

Al Mario, per aparéixer en la meua vida quan ningú ho esperava. Per posar sol als meus dies grisos, per recolçar-me en els moments difícils i per fer increïbles els dies millors.

A la meua familia, poblana i lliriana. Perquè no tinc germans però els meus cosins els sent com germans. Perquè els meus nebots, Martina, Carlos i David, són llum. Per els meus tios, Amparo, Miguel, Vicent i Enca que em cuiden com si fora una filla seua. A Rebeca, per cuidar tan bé a mon pare i per ser tan bonica. A la meua abuela (wela) Amparo, que sense cap estudi ens ha ensenyat a tots els nets les coses més importants de la vida: el valor de la familia, cuidar a tots els que estan prop, estimar-nos, ser treballadors, ser bones persones i no fer mal a ningú. Sou el meu orgull, la meua força, el meu recolzament. Vos estime amb tot el meu cor.

A ma mare i mon pare. Perquè és difícil entendre com dues persones tan diferents poden ser els meus pares i complementar-se tan bé. Perquè m'haveu demostrat dia a dia que l'amor més gran és el que es te per un fill. Perquè m'haveu ensenyat a ser lliure, valenta, forta, ambiciosa, bona persona. Perquè ho he après tot de vosaltres. Perquè mai podré tornar-vos tot el que m'haveu donat, però ho intentaré. Vos estime infinit.

Abstract

Mycoplasmas are a group of bacteria characterized by minimized genomes, limited biosynthetic capacities, and simplified metabolic networks. In this work, we explored how Mycoplasma species can be exploited for biotechnological and therapeutic applications. In chapter two, we engineered *M. pneumoniae* transposons so that they could efficiently transform different Mycoplasma species. This fact allows the generation of essentiality studies in these species, a critical analysis to identify genes that could be deleted to create attenuated vaccination strains. Chapter three developed a new method that uses high-resolution transposon and proteomics data to infer active metabolic pathways within a cell at a specific moment. This information can be used when engineering attenuated strains. In chapter four, we explored the capacity of *M. pneumoniae* to express functional human biologics *in vitro* and in mice lungs. Chapter five used the protein design algorithm FoldX and ModelX to mutate a human interleukin to enhance its properties in terms of affinity to its receptors and increased bacterial expression. Chapter six identified the secretion signals and designed synthetic promoters in the recent-discovered fast-growing *Mycoplasma feriruminatoris*.

Altogether this thesis develops the tools for exploiting Mycoplasmas for biotechnological purposes (*M. agalactiae, M. feriruminatoris*) and validates *M. pneumoniae* for human lung therapy immunomodulation.

Resum

Els Micoplasmes són un grup de bacteris caracteritzats per tenir genomes mínims, capacitats biosintètiques limitades i una xarxa metabòlica simplificada. Aquesta tesi explora com diferents espècies de Micoplasma poden ser explotades biotecnològicament o a l'àmbit de la biomedicina. Al capítol dos, es proposen transposons inicialment emprats per al bacteri *Mycoplasma pneumoniae* que poden transformar eficientment diferents soques de Micoplasma. Aquest estudi ha permés generar estudis d'essencialitat en aquestes soques i un anàlisi crític posterior dels resultats permetrà generar a soques atenuades per a vacunació. Al capítol tres es desenvolupa una nova metodologia que permet integrar dades de transposició a alta resolució amb estudis quantitatius de proteòmica per esbrinar quins són els fluxos metabòlics actius dintre de la cèl·lula en un moment determinat. Aquesta informació es pot emprar per a atenuar soques bacterianes. Al capítol quatre explorem la capacitat del bacteri pulmonar *M. pneumoniae* d'expressar funcionalment biològics humans en assajos *in vitro* i en pulmons de ratolí. Al capítol cinc es fa servir el programa de disseny de proteïnes FoldX i ModelX per mutar una citocina humana per augmentar les seues propietats inherents tant d'afinitat al receptor com en termes d'expressió bacteriana. El capítol sis identifica senyals de secreció i proposa nous promotors sintètics funcionals per al bacteri descobert recentment *Mycoplasma feriruminatoris*. En conjunt, aquesta tesi desenvolupa les ferramentes que permeten explotar diferents soques de Micoplasma per a finalitats biotecnològiques (*M. agalactiae*, *M. feriruminatoris*) i valida l'ús de *M. pneumoniae* per a immunomodulació en teràpia pulmonar humana.

Contents

[redacted]

Thesis Objectives

As an overarching objective, I wanted to develop the tools to use Mycoplasma species as a cell host for biotechnological applications, especially human therapy.

a) To design-build and test transposon plasmids that will transform a vast majority of Mycoplasmas efficiently. These transposons can allow high throughput essentiality studies to identify genes that can be deleted to create attenuated species for vaccination.

b) To develop a new methodology to ascertain active metabolic pathways within a cell by using gene transposon essentiality data and proteomics

c) To boost the use of *M. pneumoniae* as a chassis to treat human lung diseases related to immune dysregulation (i.e., inflammation).

d) To *in silico* design biologics to treat immunological disorders and optimize their expression in bacteria.

e) To identify and characterize a Mycoplasma species as a high-throughput screening system to analyze protein mutants.

List of Figures

[redacted]

[redacted]

List of Tables

[redacted]

Preface

The paradigms of science have considerably changed in the XXI century. We are now living in the era of big data and artificial intelligence (AI), where asking the right questions and hypotheses could lead to exciting results and even more stimulating scientific discussions.

One of the most appealing questions in this new era is: when all possible data types are collected and analyzed, could we fully understand how an organism works in a quantitative and predictive manner? Can we fully understand a cellular system completely? These questions have been addressed for more than 15 years in Professor Luis Serrano's laboratory using a small bacterium termed *Mycoplasma pneumoniae*. This bacterium is one of the smallest organisms capable of sustaining autonomous life and is a model organism for Systems Biology.

In recent years a vast amount of data was collected, analyzed, and discussed for *M. pneumoniae* and other Mycoplasmas *(e.g., Mycoplasma genitallium, Mycoplasma capricollum, Mycoplasma mycoides).* We have also seen the first whole-cell model that was developed for *M. genitallium.* The model was a tour de force but failed at predicting the functional effect of many gene knockouts (KO) and overexpression (OE). One explanation for this failure was that the authors used mainly data from *Escherichia coli*. However, when a similar model was generated for *M. pneumoniae* using extensive comprehensive data from this organism in our group, it also failed. These examples indicate that we are still missing critical information to understand how a cell works fully. However, we have enough information to try to engineer them for therapeutic applications.

Mycoplasma species can be tremendously valuable for biotechnological applications due to their simplicity and lack of cell walls. However, one of the problems in using these organisms is the lack of genome engineering tools, aside from transposons found to work only for some Mycoplasma species. This thesis partly addresses this limitation by developing a universal transposon vector to modify their genomes. Thanks to the development of this universal transposon for Mycoplasmas, we were able to identify and quantify the active metabolic pathways in some of these species using genomics and proteomics data. This methodology could generate helpful information when designing defined culture media for their growth or using these microorganisms as expression platforms.

[redacted]

1. Chapter 1. Introduction

1.1 To know History is to know life

The biotechnology era (i.e., the usage of living systems to generate various products) started thousands of years ago. The ancient History of biotechnology started through plant and animal domestication and selective plant breeding. Also, the usage of yeast species to ferment fruits and grains is dated about 2500 BC. Many years later, in 1665, Robert Hooke presented the first illustration of a microorganism. In the following years, Antoni van Leeuwenhoek described protozoa and bacteria. Thanks to the technological development of image magnification devices technological development (Adamu et al. 2016), these advances were possible. Two centuries after, in the mid-1880s, the medical doctor Theodor Escherich observed a type of bacteria (later described as *Escherichia coli*) in both the intestine of healthy children and those affected by diarrhea. This observation was the first evidence of human body colonization by a vast diversity of microorganisms in healthy states, latterly termed the human microbiome.

The first application of microorganisms to treat or prevent human diseases was vaccination (1796, Edward Jenner, and the smallpox vaccine). Just after that, in a humanitarian mission lead by Francisco J. Balmis, children infected with smallpox were transferred by boat from Spain to Mexico to immunize the population and save thousands of lives. Years later, Alexander Fleming, in 1928, discovered that some fungal species could produce an anti-bactericide product, penicillin, which was the first antibiotic described (Fleming, 1929). Approximately 50 years after Fleming's discovery, genome modification's first techniques appeared when Herbert Boyer and Stanley Cohen transferred genes from one organism to another (Cohen 1923). A few years after, the synthesis of novel antibiotics for chemotherapeutics was developed by Paul Enrich (Ehrlich, 1960).

In 1995, DNA sequencing technology's development achieved a milestone for humanity: the sequencing for the first time of an organism's genome, *Haemophilus influenzae* (Fleischmann et al., 1995). In the following years, the cheapening of sequencing techniques has made organism sequencing accessible for many laboratories. As a result, we have thousands of genomes from a vast range of organisms, from bacteria to humans, transforming the fields of medicine and biotechnology.

The possibility of doing massive sequencing of many organisms simultaneously (metagenomics) has favored the launching of The Human Microbiome Project (started in 2007), which has shown stable microbial communities in different parts of the human body (see Figure 1). Altogether, the study and identification of tissue-specific microbial populations in health and disease have paved the way to revolutionize medicine.

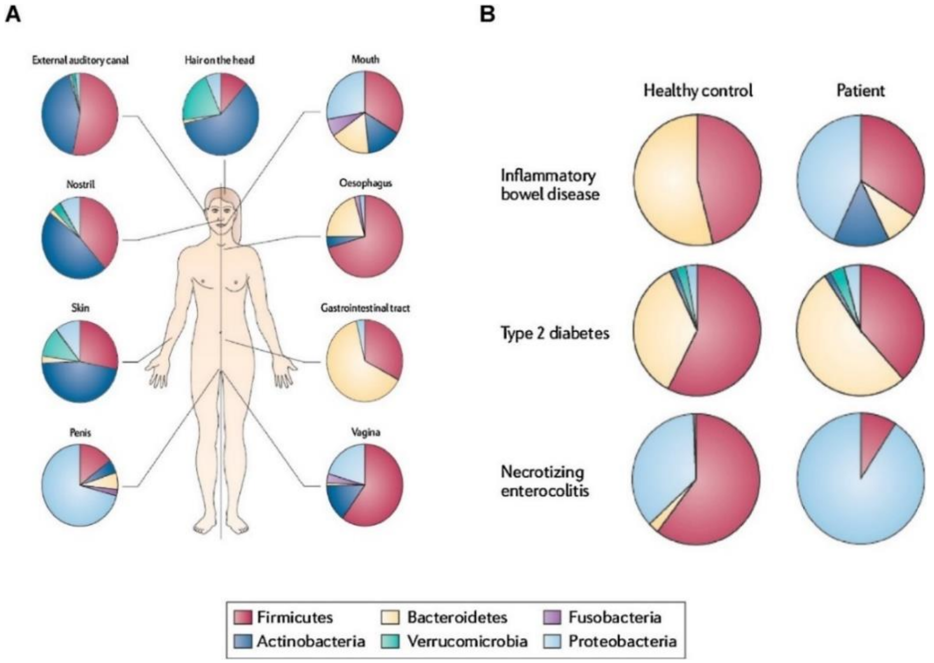


Figure 1. The human microbiome aims to characterize the microorganisms populations that co-habit in a healthy human body. The bacterial communities vary

across the tissues (A). Then, dysregulations in the microbiota in distinct tissues might impact the appearance or aggravation of different diseases such as diabetes (B). The figure is taken from (Blum 2017).

1.2 Synthetic Biology

The advances in DNA synthesis, molecular cloning, and DNA sequencing technologies have opened the field to re-design existing molecules, entities, networks, and organisms to repurpose their existing properties or introduce new functionalities. The exercise of engineering (the use of science and maths to solve problems) biological entities to create a variation of the existing forms or introduce novel functions is termed Synthetic Biology. This definition also applies to converting non-natural or biologically non-existing entities for biological purposes (Synthetic Biology and Biodiversity European Commission, Issue 15 September 2016).

There are classic and well-known examples of Synthetic Biology exercises. For instance, the yeast *Saccharomyces cerevisiae*'s metabolic engineering for the production of artemisinic acid, the precursor of the anti-malarian drug artemisinin (Ro et al., 2006). There are also other less-known cases with great potential: i.e., i) the metabolic re-design of the soil bacterium *Pseudomonas putida* towards a linear glucose catabolism pathway with enhanced carotenoid production properties (Sánchez-Pascuala et al., 2019); or the implementation of gene networks into bacteria for understanding fundamental biology paradigms such as the action mechanisms of antibiotics that have evolved into the generation of components to overcome antibiotic resistance (Laboratory of Prof. JJ Collins, Wyss Institute, MIT, US).

If Synthetic Biology aims to modify biological entities the same way engineers do, defining standards and pieces to be reused and mixed is critical. In this direction, different repositories have been established in the past years (iGEM Repository, SynBioHub, JBEI-ICE) to store genetic parts (promoters, transcription factors, terminators, plasmids, DNAs), also termed Biobricks.

The assumption behind the repository is that biologicals are expected to behave similarly, independently of the context when used in the same species (principle of standard). However, when these biological parts are transplanted even to a different strain of the same species or in the context of other introduced genetic circuits, they do not always behave as expected. This problem is defined as a lack of orthogonality (Vilanova et al., 2015). Thus, it is critical to specify the system's genetic details and the experimental information (conditions, mathematical models) when doing Synthetic Biology exercises. In this line, Synthetic Biology Open Language (SBOL) also includes information about the designs' history more abstractly. Likewise, the description of units should be addressed to guarantee interoperability. In this respect, recently, the Synthetic Biology community has united in the EU-funded project BioRoboost to generate a standardized and flexible biological toolbox catalog (https://standardsinsynbio.eu/, visited on 13/01/2021). The project aims to identify the weakest points in the standardization procedure and help develop and define different bacterial chassis (see chapter below for a definition) for specific applications (i.e., production of biologicals *in vitro*, delivery of molecules in the gut, bioremediation, etcetera). We could reuse and mix various gene circuits in a particular chassis with more success chances by having these standard chassis.

1.3 Bacteria and their applications in human health. The concept of chassis.

In the early 1930s, the term chassis was adopted to describe a car's structural frame to which car components are added. This term, chassis, is nowadays widely used in Synthetic Biology. A chassis in Synthetic Biology is a biological entity where distinct components or functionalities can be added or removed in consonance with a specific purpose.

There is no universal concept of chassis that will vary depending on the applications. For example, if we need a bacterium that will grow cheaply and produce grams of proteins *in vitro*, *E. coli* will probably be the organism chosen. Nevertheless, if we would like to deliver a therapeutic protein in the human colon, we could use a bacteria that is non-pathogenic, safe, and that survives the transit through the digestive system, i.e., *Lactococcus lactis*. Furthermore, within the same organism, we could have different chassis depending on the application and the organ to be treated.

Simultaneously to the Synthetic Biology field explosion, microorganisms have emerged as cellular systems that can be exploited as therapeutic payloads, drug carriers, or biofactories. The introduction of microorganisms in medicine (bacterial therapies) has refined the concept of chassis. Therefore, novel functionalities are plugged into existing bacterial backgrounds, covering diverse applications such as treating various infectious diseases, cancer, or locally producing interesting molecules at physiological levels.

The two critical aspects that make bacterial therapies attractive are the continuous delivery of therapeutic molecules and the possibility of local delivery, thus overcoming the limiting toxicological features that appear when drugs are administered systematically. Finally, on more economic terms, it reduced the industrial production chain's price of biologicals for therapy by joining in one organism the production, purification, and administration steps.

The set of bacterial species explored for therapeutic strategies comprises well-known species such as *E. coli*, part of de human-gut microbiome, *L. lactis,* or *Salmonella enterica* subsp. *enterica serovar Typhimurium*. This last species is a human pathogen exhibiting preferential growth in human tumors over healthy tissues. An attenuated version of this bacterium (VNP-2009) expressing cytosine deaminase enzyme has been exploited for its capacity to convert the prodrug into a drug (i.e., 5-fluorouracil) in tumor core (Nemunaitis et al., 2003). Likewise, engineered versions of the pathogenic bacteria *Listeria monocytogenes*, a facultative anaerobic bacterium responsible for the listeriotic infection, have been tested in clinics for producing tumor-antigens and stimulate immune responses. Also, an attenuated chassis based on *L. monocytogenes* has been engineered to express epitopes that exhibited aberrant expression in ovarian and pancreatic adenocarcinomas or Human Papilloma Virus (HPV) derived cancers (Friedman et al., 2000).

Lactococcus is a genus deeply explored in biology, and various members of this family are considered safe (GRAS) (e.g., *L. lactis* or *Lactococcus plantarum).* These species are non-pathogenic, non-invasive and universally used for biotechnological purposes and diet supplementation. Organisms of this genus are used for the treatment of gut, mouth, and vaginal diseases. The strategies used include *Lactococcus gasseri* engineered to deliver glucagon-like peptide 1 (GLP-1) for reprogramming intestinal gut cells to produce insulin in response to glucose (Suzuki et al., 2003). The diabetes mellitus type-I treatment using *L. lactis* as a delivery system combining proinsulin and interleukin-10 (IL-10) showed decreased β-cell disruption and improved cell regeneration. IL-10 is an anti-inflammatory cytokine, but its systemic administration at high-doses might induce pro-inflammatory interleukins such as interferon-γ (IFN- γ) with toxic effects (Tilg et al., 2002). *L. lactis* mediated delivery of IL-10 showed a functional response similar to that obtained with a systemic 10.000 times higher dose of recombinant protein, decreasing significantly undesired side effects (Steidler et al., 2000). The same bacterium was used to study the therapeutic agent's permeability. *L. lactis* expressing IL-10 was detected in the inflamed intestinal mucosa, providing evidence for the mucosa intake of *L. lactis* in inflamed tissue, therefore bettering the treatment efficiency (Waeytens et al., 2008).

Another exemplification of live therapeutics' superior capabilities of *L. lactis* compared to recombinant protein delivery is interleukin 27 (IL-27). In a mice model of intestinal colitis, the IL-27 delivered by *L. lactis* protected the mice from enterocolitis, modified the CD4$^+$ population and IL-17$^+$ cells in the gut-associated tissue, and boosted IL-10 T-cells production. This study also supports the idea that bacterial protein *in vivo* delivery outperforms systemic administration of recombinant IL-27 to reduce colitis phenotype (Hanson et al., 2014).

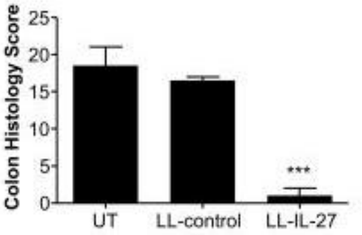


Figure 2. Colon histology Score in an intestinal model of enterocolitis for untreated mice (UT), L. lactis non-expressing cells (LL-control), or L. lactis expressing IL-27 (LL-IL-27). The figure is taken from (Hanson et al., 2014).

In this context, different biocontainment strategies should guarantee the prevention of the engineered bacteria escaping or disseminating through the body. The scenarios explored to create attenuated strains are the introduction of tuneable genetic circuits for programmed senescence, the generation of conditional auxotrophic organisms, or the generation of attenuated strains with no replication capacities *in vivo*. One example is the replacement in *L. lactis* of native essential *thyA* for human *il10.* Consequently, this strain's survival and growth are rigidly dependent on thymidine or thymine (Steidler et al., 2000).

However, bacterial expression of biomolecules is faced with some issues related to the organism's possible pathogenicity. Gram-positive and gram-negative bacteria have cell walls that contain many pro-inflammatory molecules and make it difficult in some cases to secrete proteins. Many of the most exciting biomolecules have many disulfide bridges and require complex chaperon systems or glycosylation to fold correctly. Thus, it is challenging to express these proteins in bacteria, where glycosylation patterns are different, and the enzymes and chaperones involved in the folding and formation of disulfide bridges are only found in some cases in the periplasmic space.

Canonical antibodies (Abs) are large heterotetrameric protein complexes with two heavy chains and two small polypeptides light chains. These chains fold in functional units (Ig domains) and adopt a Y shape. The variable part is responsible for antigen recognition properties. The fragment crystallizable (Fc) has effector functions when bonded to Fc receptors (cell-mediated cytotoxicity, cell-mediated phagocytosis) and has been tuned to boost or disable effector functions in therapy. The production of antibodies is often not compatible with microbial expression. The different subunits must cluster in the space and ensemble in a suitable conformation.

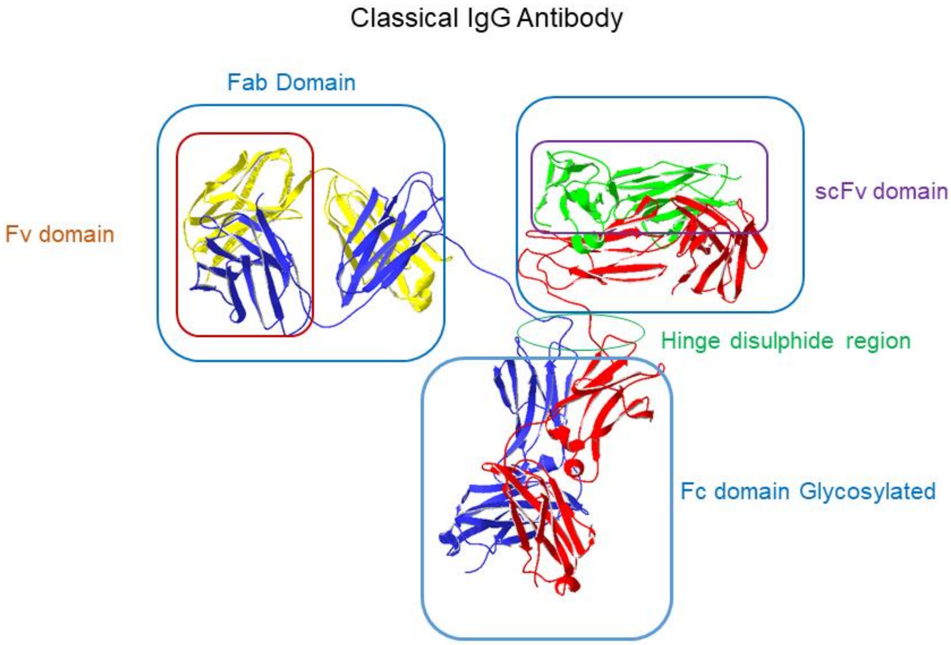


Figure 3. Structure of an IgG antibody and different engineered antibody versions. In blue, Fc domain with glycosylations. In green, the hinge regions with three disulfide bridges. In purple, engineered version, single-chain fragment variable (scFv). In light blue, Fab domain. Overall, the classical representation of an IgG in a protein 3D viewer.

Additionally, they are glycosylated on the Fc region, and this modification is essential for the Fc to bind their receptors (FcγRs) and therefore perform an effector function (Kang and Jung, 2019). Finally, they have multiple disulfide-bridges (i.e., three) at the hinge region that stabilize the protein structure and are difficult to be made in the correct conformation outside of a mammalian or insect cell.

Despite this hurdles there are reports of mutated aglycosylated IgG variants expressed in the bacterial host that bind to FcR neonatal and increase their half-life in the serum (Jung et al., 2010). Also, full-length antibodies have been expressed in the cytoplasm of engineered *E. coli* that allows the formation of disulfide bridges into the cytoplasm (Robinson et al., 2015).

However, these cases so far have been reduced to anecdotal cases, and antibodies are still produced in mammalian and/or insect-producing systems. Nature has found a way to simplify the complexity of full-length antibodies in the *Camelidae* family: heavy chain-only antibodies known as VHHs or nanobodies. Nanobodies are simplified versions with no light chain that retain full antigen-binding capacities, are soluble, robust to chemical and temperature changes, and are expressed in bacteria (Muyldermans, 2013). Their potential use in live therapeutics clinics has been shown using *L. lactis* for colitis expressing anti-TNF nanobodies (Vandenbroucke et al., 2010). Besides, nanobodies have been projected as ensembling modules for a set of well-characterized microbial communities to treat microbiome dysbiotic diseases (de Lorenzo, 2008; Timmis et al., 2018). Therefore, nanobodies have the potential to overcome some of the folding limitations of full-length antibodies due to their simplicity. However, given that the number of available canonical antibodies is higher than nanobodies, there is a clear need to generate novel engineered variants that can be cheaply expressed in a microbial host for human therapy.

1.4 Mycoplasma species as chassis in Synthetic Biology

The famous quote by Nobel winner Richard Feynman' What I cannot create, I do not understand', has become a sort of life's motto for some synthetic biologists. Following Feynman's bottom-up approach, Mycoplasmas, proposed as the minimal self-replicating form of life, are appealing chassis to investigate mechanistically how a cell works and re-construct the minimal set of components required to sustain life.

Mycoplasmas are part of the *Mollicutes* class, evolved from gram-positive around 600 million years ago through gene minimization, termed degenerative evolution, from a gram-positive bacteria branch (Woese et al., 1980). They can colonize a wide range of organisms, such as plants, arthropods, or many vertebrate animals (e.g., chicken, sheep, goats, pigs, mice, or lambs) and humans. Mycoplasma is the best-known genus of *Mollicutes* typified by the lack of a cell wall, low GC content and low degree of redundant metabolic pathways, and limited biosynthetic capacities. The Mycoplasma genus has been classically considered models for studying basic life principles such as transcription and translation. Indeed, three landmarks of modern biology have been achieved using Mycoplasma species as model systems (i.e., *M. pneumoniae*, *Mycoplasma mycoides*, and *Mycoplasma genitalium*).

i)M. pneumoniae was the first organism for which the whole proteome, transcriptome, and genome were simultaneously available (Güell et al., 2009; Kühner et al., 2009; Yus et al., 2009a) and used as a model for Systems Biology to understand an independent organism from a quantitative perspective.

ii) The first chemically synthetic cell was generated from *M. mycoides*, denominated JCVI-syn3.0. This cell is organized into 473 essential genes and used a chemically synthesized minimized genome where all essential genes for life were retained and resulted in the smaller genome of any alive replicating form in nature (Hutchison et al., 2016). The version JCVI-syn3.0 has been refined after developing gene essentiality studies for JCVI-syn1.0, the first synthetically organisms generated. These studies have been performed using Tn5 transposons and helped in tuning the minimal viable genome model (JCVI-syn3.0), see Figure 4.

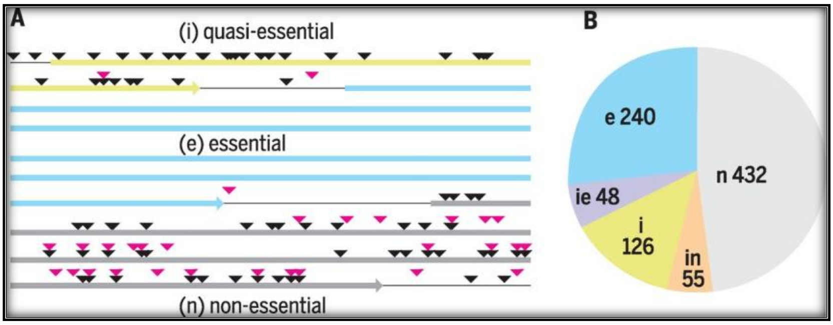


Figure 4. Essentiality study performed in synthetic bacterium JCVI-syn1 using transposition insertion technology. Genes were classified as essential (e), non-essential (n), quasi-essential (i), genes with minimal growth disadvantage (in), and those whose disruption showed severed growth effects (ie). The figure is taken from (Hutchison et al., 2016).

iii) The first whole-cell computational model (Karr et al., 2012) that succeeds in predicting some cell phenotypes was accomplished for *M. genitalium*.

However, the aim of fully understanding an organism using *Mycoplasma* species evidenced some practical challenges associated with these organisms: the lack of a chemically-defined medium and the paucity of genome engineering tools to generate targeted cell mutants.

As mentioned above, Synthetic Biology aims to engineer organisms to tackle biomedical or biotechnological problems. This thesis's objective goes in the direction of paving the way to use *Mycoplasma* species for a set of applications.

1.5 Mycoplasma species as pathogens

Mycoplasma species infect plants and animals, and there is hardly any animal species that does not have an associated Mycoplasma species. Infection by Mycoplasma of domesticated animals causes severe economic damage. There are 23 *Mycoplasma* species identified that infect poultry. Among them, *Mycoplasma gallisepticum*, *Mycoplasma synoviae*, *Mycoplasma meleagridis*, and *Mycoplasma iowae* cause dramatic economic losses in the poultry industry. On the other hand, cattle animals are affected by several species such as *Mycoplasma hyopneumoniae, Mycoplasma agalactiae,* or *Mycoplasma bovis*. The range of diseases caused by their infection is extensive, infecting mammary glands, the respiratory tract, or the reproductive system. As an illustration, the swine industry's two main significant challenges are infections caused *by Mycoplasma hyopneumoniae* or PRRSV virus. In the EEUU, the percentage of animals affected by *M. hyopneumoniae* is 17.6 in the breeding period, 10.0 in the nursery period, and 34.3 in the finishing period (https://www.pig333.com/articles/economic-impact-of-mycoplasma-hyopneumoniae-on-pig-farms_8936/, visited September 2020)

Most of these infections are difficult to treat with antibiotics due to the lack of cell wall, and there is a need to develop novel vaccines against these animal pathogens. Although some vaccines are already available (Kanci et al., 2018; Zhang et al., 2014), most of them rely on naturally occurring attenuated strains, and as reported for *M. hyopneumoniae* by the pharma company Merck Sharp & Dohme (MSD), they are not fully effective. Synthetic Biology opens the door to the rational engineering of microorganisms and can be applied for vaccine design based on removing the main pathogenic determinants of a bacteria but allowing it still to stimulate an immune response.

The engineering of a bacterial chassis for vaccination would ideally rely on detailed knowledge on which parts of the genome are essential, fitness, or non-essential under the conditions it will be used. Also, we need to establish a gene-function relationship and ascertain which genes are involved in an organism's pathogenesis, followed by removing those genes involved in pathogenicity while keeping those necessary to keep the bacteria alive in the injected animal to maximize the response. This information can be obtained through the combination of generating saturated libraries of transposon mutants plus tracking the insertions through ultra sequencing across different passages (HITS) (Figure 5). This allows classifying the genes in essential (E), fitness (F), and non-essential (NE) categories. The E genes are those genes required for sustaining life under the conditions analyzed, while those that can be deleted without compromising survival are categorized as NE. Genes whose disruption might cause a growth impairment are classified as F.

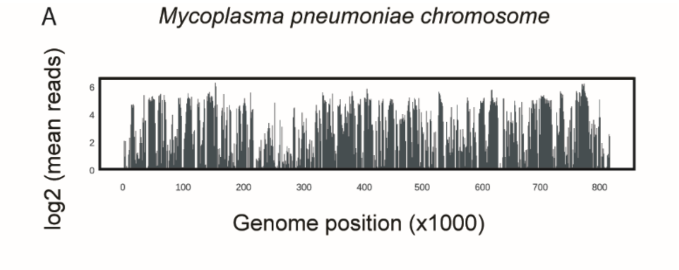


Figure 5. Representation of the transposon reads mapped across the genome of M. pneumoniae. In Y-axis is represented in a min (0) to max (6) scale the log2 of mean reads detected at each position (X-axis). The figure is taken from (Montero-Blay et al., 2019).

For essentiality studies, transposons derived from *Staphylococcus aureus*, Tn1004, have been widely used in the Mycoplasma community. Aside from having random insertions with no significant bias for particular sequences, it is essential to ensure that many chromosome positions in the population have transposon insertions (high-coverage) to have reliable statistical data. E genes for life will not have transposon insertions because their disruption prevents the cell from surviving.

When this thesis was started, there were few essentiality studies for Mycoplasmas species (i.e., *M. genitalium* (Glass et al., 2006), *M. pneumoniae* (Lluch-Senar et al., 2015), *Mycoplasma pulmonis* (Dybvig et al., 2010), *M. bovis* (Sharma et al., 2014) and *Mesoplasma florum* (Baby et al., 2018). However, some of these studies had very little coverage (e.g., coverage for *M. bovis* essentiality study is 0.03% and 0.56% for *M. genitalium*), which decreased confidence in the essentiality assignment, especially for small genes.

As indicated above, there is a clear need for vaccines against Mycoplasma species, especially in veterinary medicine, to reduce the number of antibiotics supplied to the animals to prevent infections derived from these microorganisms. For this, we need to perform essentiality studies of the target species with high reproducibility and coverage, which in turn requires efficient universal *Mycoplasma* transposons. This is addressed in the second chapter of this thesis.

1.6 The metabolism of Mycoplasmas

One of the most critical requirements in the therapeutics manufacturing industry is reproducibility and lot-to-lot homogeneity. When applied to industrial microbiology, it is imperative to define the exact composition of the growth media required to grow certain microorganisms. Historically, this task has been challenging for Mycoplasmas, where their simple metabolisms force the bacterium to acquire the vast majority of the metabolites from the medium. The media generally used for their growth (e.g., SP4, Hayflick) contain animal-derived serum, which is considered a risk factor because it can introduce contaminants as viruses and result in low reproducibility changing batches. Therefore, the development of serum-free media for growing Mycoplasma species for vaccination or therapy is a must.

For *M. pneumoniae*, there is a serum-free medium containing 26 components (Yus et al., 2009b) that also allowed the growth of *M. genitalium*. However, growth was not very robust with this medium and prevented high-biomass culture media of *M. pneumoniae*. A new version of serum-free media tested successfully on a larger scale has now been developed in the laboratory by Raul Burgos, identifying experimentally a key component (sphingomyelin) missing in the medium of Yus et al., (Gaspari et al., 2020; Yus et al., 2009b). No serum-free medium has been published for other relevant species, such as *M. bovis* or *M. hyopneumoniae*. We have evidence in the group that the defined medium for *M. pneumoniae* does not work on *M. hyopneumoniae* without modification of concentrations and components (*Burgos et al. in preparation*). In a detailed metabolic analysis of two Mycoplasma species (*M. bovis* and *M. gallisepticum* (Masukagami et al., 2017)), it was shown that both species significantly differ in the use of carbon sources.

To make a defined medium for a particular Mycoplasma species, we need first to build a comprehensive metabolic map and determine the essential and optional metabolic pathways used for each microorganism and its fluxes' directionalities as was done for *M. pneumoniae*.

There are various approaches to measure, quantify, and represent cellular metabolism through mathematical models. For example, Genome-Scale metabolic models (GSMM) can automatically reconstruct a full metabolic map. However, this requires later a maps' curation process, reviewing each gene, metabolite, and reaction to avoid futile loops or unconnected reaction steps (i.e., gap-filling). The performance of these *in silico* tools has been assessed based on a set of features important in GSMM (Mendoza et al., 2019) for two different bacteria: *Lactobacillus plantarum* (Teusink et al., 2006) and *Bordetella pertussis* (Branco Dos Santos et al., 2017). In the analysis, none of the *in silico* methodologies results were outstanding in all categories indicating that for each case, the tool utilized should be selected *ad hoc* for the purpose.

In the same way that methodologies were developed to connect networks, computational tools to fill and curate those metabolic gaps have been proposed. These methods are supported by a database of enzymes, metabolites, reactions, and directionalities, and the objective is to connect in a logical manner all reactions proposed. The existing methodologies are limited to a set of known reactions contained in the backup database. In a systematic analysis of the existing *in silico* methodologies for gap filling (Karp et al., 2018), one of the main conclusions of the extensive analysis is that the larger the set of reactions, the larger the number of unneeded reactions suggesting that gap fillers should look for results containing minimal reactions. Finally, we should mention that all these approaches rely on the sequence homology with enzymes of well-characterized organisms like *E. coli* to assign functionality. This could be dangerous since only one mutation is enough to convert a lactate dehydrogenase into a malate dehydrogenase (Boucher et al., 2014), for example. In Mycoplasmas, MPN483 is a glycosylase in *M. pneumoniae* uses UDP-galactose preferentially (Klement et al., 2007), while in the closest related *M genitalium* species, its MG517 orthologue prefers UDP-glucose (Andrés et al., 2011). Thus sequence conservation and homology do not always indicate that the two orthologue enzymes will prefer the same substrate and result in the same product. Moreover, functionalities assigned to classical enzymes like Pyruvate kinase in glycolysis often hide other functions like nucleotide triphosphates' synthesis (i.e., in Mycoplasma species (Pollack et al., 2002)). As a result, we could observe enzyme gaps in a metabolic map like the absence of a nucleoside diphosphate kinase in *M. pneumoniae* that is covered by pyruvate kinase.

The simplicity of Mycoplasmas species with around 700 genes allows the manual building of their metabolic maps. One of the first Mycoplasma species for which a complete metabolic map was annotated, and as mentioned above, based on it, a defined medium was produced for *M. pneumoniae* (Yus et al., 2009b). Not only was the map generated, but also the experimental measurement of fluxes and metabolites (Maier et al., 2013) was performed, and a metabolic flux balance analysis (FBA) was done (Wodke et al., 2014) (Figure 6). Some metabolic aspects of other Mycoplasma species (i.e., *M. gallisepticum* and *M. bovis*) have been characterized by isotope tracking (Masukagami et al., 2017). There is also a recent metabolic map available for *M. hyopneumoniae* (Kamminga et al., 2017).

However, as indicated above, not all metabolic map pathways annotated in a metabolic map are used simultaneously, and the reversibility of reactions is not always exact. Even for the simplest organisms, multiple pathways can be used to produce a metabolite. For instance, in theory, *M. pneumoniae* can use ascorbate, glycerol, glycerol-3P, glucose, fructose, mannitol, and phosphatidylcholine as carbon sources. However, in reality, it grows very well in glucose and mannose, poorly in glycerol and phosphatidylcholine, and almost not with the rest of the carbon sources (Yus et al., 2009b).

Thus to understand which pathways will be preferentially used in a specific condition, we need methods that can predict the dynamics of metabolism under different circumstances.

Flux balance analysis (FBA) aims to quantify the various flows in a metabolic network. These models consider each reaction's stoichiometry, the kinetics of each enzyme participating in it, and the metabolites consumed or produced in each metabolic reaction. This approximation uses linear optimization to achieve the steady-state flux distribution and maximizes a set objective (e.g., growth rate, ATP, protein production). One of the problems of classical FBA models is that they did not integrate enzyme abundance, and therefore all pathways are considered functional.

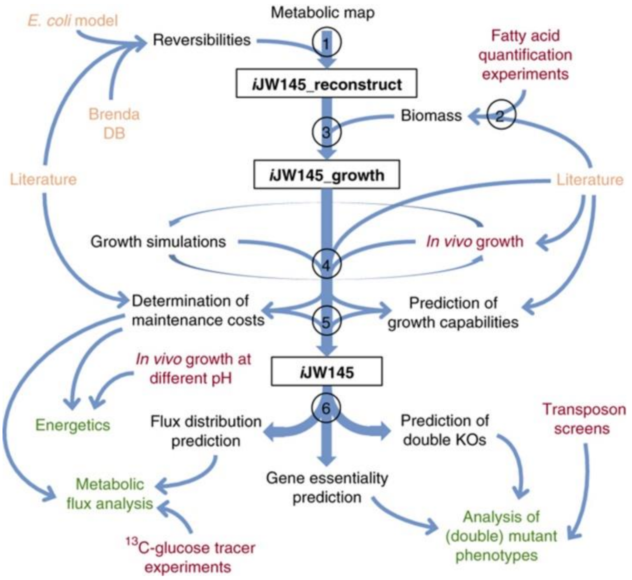

E. coli model
Metabolic map
Reversibilities
1
Fatty acid quantification experiments
iJW145_reconstruct
Brenda DB
Biomass
2
3
iJW145_growth
Literature
Literature
Growth simulations
In vivo growth
4
Determination of maintenance costs
5
Prediction of growth capabilities
In vivo growth at different pH
iJW145
Transposon screens
6
Energetics
Flux distribution prediction
Prediction of double KOs
Metabolic flux analysis
Gene essentiality prediction
Analysis of (double) mutant phenotypes
^{13}C-glucose tracer experiments

Figure 6. Integrated set of experimental metabolic information (red) plus bibliographic data (orange) for the generation of a model for M. pneumoniae's metabolism that reproduces fluxes in silico and predicts energetics, FBA, and analyses mutant phenotypes (green). The figure is taken from (Wodke et al., 2014) and based on (Yus et al., 2009b) metabolic network reconstruction.

The latest versions of FBAs integrate gene expression profiling to solve this issue. The method tries to find a flux distribution that fits with environment-specific gene-expression data assuming that gene expression changes correlate with flux values, which might be questionable. This methodology demonstrated a good correlation of flux predictions with experiments using ^{13}C labeling in *Bacillus subtillis* (Thanamit et al., 2020) or *E. coli* (Pandey et al., 2019). For one of the organisms for which more datasets are available, *E. coli* multi-omics data at different layers has been integrated (transcriptome, proteome profiles, metabolomes, fluxome) from more than 600 different experimental conditions. This Multi-Omics-Model and Analytics (MOMA) showed acceptable performance in predicting cellular growth and demonstrated a robust capacity to estimate metabolite concentration despite the high degree of variability in the data (Kim et al., 2016; Macklin et al., 2020).

Despite all the methodologies proposed to ascertain the intricate and interconnected set of active metabolic reactions, there is still a clear need to determine each organism's pathways across different conditions in a high-throughput manner. Also, since metabolic enzymes might have more than one metabolic functionality in the cell (i.e., pyruvate kinase mentioned above), other types of available omics data besides genomics or transcriptomics can be interrogated to verify if a given enzyme participates in another metabolic pathway. In chapter three, this thesis shows how gene essentiality and protein abundance could help determine active metabolic pathways and directionality.

1.7 *M. pneumoniae* as a living pill to treat human diseases

M. pneumoniae has a tremendous potential of being used as chassis in biomedicine to treat pulmonary diseases. The lung is so far an orphan organ for bacterial therapies even though lung diseases are the leading cause of human death (Forum of International Respiratory Societies and European Respiratory Society, 2017). Plus, until recently, the lung's microbiome was not characterized and thought to be sterile (Huffnagle et al., 2017).

The advantages of this bacterium from a Synthetic Biology point of view are: i) its small genome (816 kbp) translated in simplified genomic and metabolic networks that diminish the risk of unpredicted interaction with the host or the engineered circuits, ii) the lack of cell-wall (lipopolysaccharides-free) and therefore can be delivered in combination with antibiotics targeting the cell wall, iii) its mild- pathogenic behavior in the lung, iv) the described virulence factors in the literature (Becker et al., 2015; Chaudhry et al., 2016; Somarajan et al., 2010) v) its genome codes an inherent mechanism of biocontainment, by using the stop codon UGA to encode for tryptophan (Andachi et al., 1987) vi) low recombination efficacy and, therefore, a low rate of horizontal transfer.

Until recently, the lack of targeted genome engineering tools was one of the main drawbacks of using *M. pneumoniae* for clinics. This has changed very recently, and we now have a toolkit to generate targeted knock-in and knock-out mutants (Piñero-Lambea et al., 2020) and biosafety circuits based on Cas9 as a counter-selector marker (unpublished data). Thanks to this, two attenuated chassis versions (Mycochassis) have been generated [redacted]

[redacted]

[redacted] It has also been demonstrated that the [redacted] chassis can dissolve biofilms *in vivo* in intradermal mice catheters using enzymes (alginate, hydrolase). The buildup of the resources needed to exploit *M. pneumoniae* as therapeutic chassis for treating microbial lung infections in ventilator-associated pneumonia (VAP) is being explored by a spin-off company incorporated in 2020, Pulmobiotics S.L.

This thesis aims to investigate this bacterium's capacity to be used in diseases with a dysregulated immune response. [redacted]

1.8 Lung diseases involving immune system dysregulation

The lung is a tissue exposed continuously to substances, toxins, dust, pollens, pathogenic microorganisms, or pollutants in the air we breathe. The lung's immune system has evolved towards an equilibrium between protecting the tissue against inhaled damaging substances and preventing an aberrant immune activation that can damage the tissue (Figure 8).

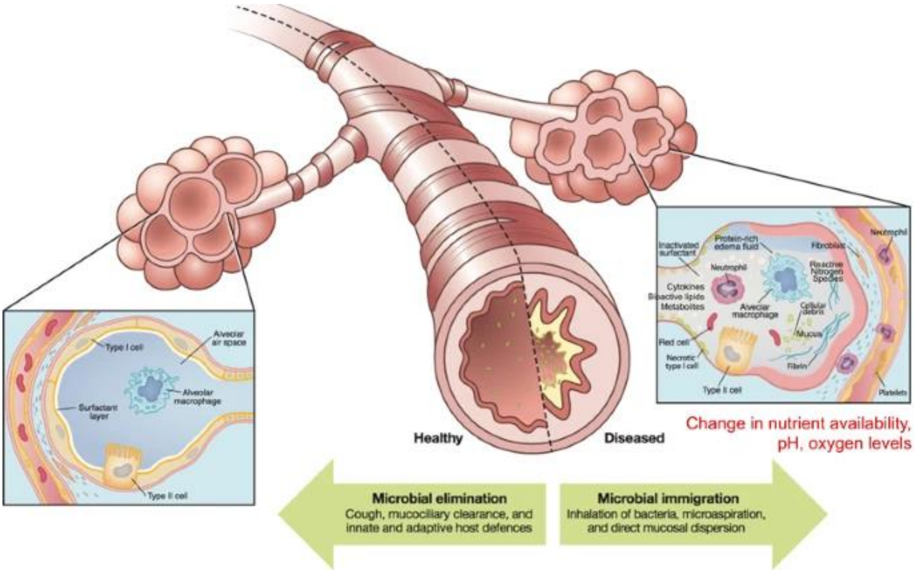


Figure 8. The scheme represents how lung architecture maintains tissue homeostasis and protection when exposed to environmental factors that modify lung microbiota. The figure is taken from Huffnagle et al., 2016.

The defense mechanism starts with the tracheobronchial tree's anatomical structure. The mucociliary escalator based on motile cilia expulses the large particles and the mucus accumulated. The mucus composition is a mixture of distinct components such as anti-proteases, antimicrobial peptides (e.g., β-defensins, cathelicidin, lysozymes, lactoferrin), antioxidants, and IgA. Additional protective molecules like lung-collectins, surfactant protein A (SP-A), and surfactant protein B (SP-B) can be found in the alveoli. Second, the lung contains cells implicated in innate lung immunity (macrophages, neutrophils, alveolar epithelial cells, and monocytes), offering the first response to inhaled particles (Riches and Martin, 2018).

Alveoli respond to microbes' presence because they recognize a group of microbial features called Pathogen Associated Molecular Patterns (PAMPs) (e.g., lipids, sugars, and cell-wall, DNA or RNA sequences). These molecules bind to receptors expressed in alveolar macrophages, epithelial cells, and neutrophils. When those PAMPs are recognized, a molecular cascade is activated, chemokines and pro-inflammatory interleukins are delivered, and genes implicated in phagocytosis, killing, and degradation are expressed.

Many different diseases that affect the human lung are distinguished by dysregulated immune system activity: in the direction of an exacerbated inflammatory response (infections, lung fibrosis, chronic obstructive pulmonary disease (COPD) or lung injury) or towards immunosuppression (i.e., lung cancer, chronic infectious diseases). In the context of inflammation, idiopathic pulmonary fibrosis is a progressive and chronic lung disease distinguished by lung scarring and histopathological signature of usual interstitial pneumonia (Barratt et al., 2018). The main consequence is the aberrant reparation of injured epithelial alveolar tissue. This disease affects more than three million people globally, and its incidence is related to aging, causing dyspnoea and cough (Martinez et al., 2017). There are treatments available for this disease that slow down its progression (N-acetylcysteine, corticosteroid pills, piperidone, nintedanib), but there is no cure.

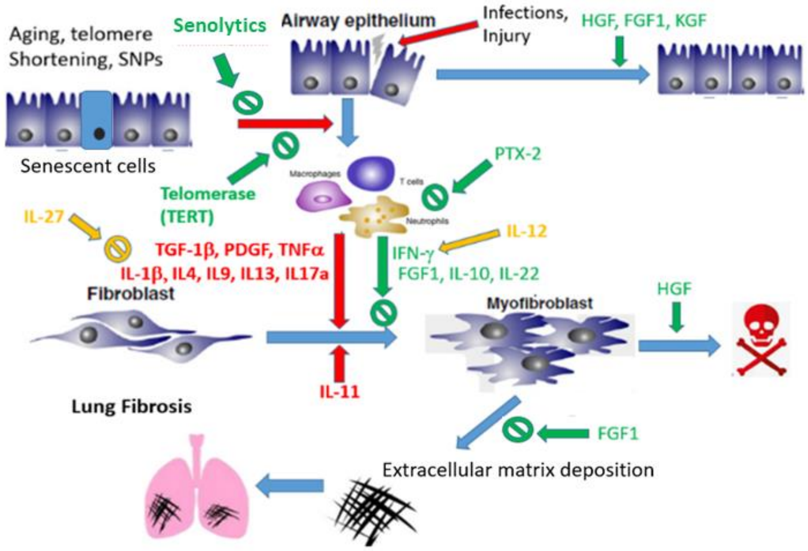


Figure 9. Schema showing the stimulatory and inhibitory actions of different proteins on IPF. Red arrows represent stimulation. Green arrows ended with a circle, inhibition; without a circle, stimulation and yellow arrows show indirect effects.

[redacted]

1.9 Protein design tools. Going beyond nature

Nature does not shape natural existing proteins towards the most optimized version for their expression, affinity of pharmacological properties. Therefore, there is room for improvement of these properties by mutagenesis. Protein engineering strategies have been used to modify immunotherapeutic biologicals. For example, IL-15 is a promising anti-tumor component needed to induce proliferation and cytotoxic activity. The molecule has been modified to bind Apo-lipoprotein I (a component of lipoproteins), whose receptors are overexpressed in a tumor cell to enhance its targeting properties

(Ochoa et al., 2017). Other examples are the conjugation of IL-10 or IL-2 to polyethyleneglycol (PEG) or the fusion of pre-conjugated IL-15 to an IL-15Rα-IgG1 Fc domain to increase protein half-life in the serum and improved antitumor activity (Zhao et al., 2019).

In 2019, a novel mimic of IL-2 and IL-15 was designed *de novo* using the protein engineering software Rosetta Suite as *in silico* engineering platform (Silva et al., 2019). In this engineering strategy, IL-2 helical fragments were assembled using a fragment-database of four residues to recapitulate the starting body's shape. Then, helices were connected using a loop fragments database, resulting in a wholly connected protein backbone, further mutating each side-chains to minimize energy. The result was a potent agonist of IL-2 and IL-15, with increased therapeutic potential tested in two different murine models of cancer (melanoma and colon) with increased stability, reduced toxicity, and imponderable immunogenicity. The software used in this above-described work, Rosetta, is an open software that includes algorithms developed by Dr. Baker (University of Washington, US) to perform computational modeling of a protein structure analysis and ultimately apply it to design novel molecules. The competitor in the field is the FoldX (Schymkowitz et al., 2005) and ModelX tool suite (Blanco et al., 2019; Cianferoni et al., 2020)

FoldX is an empirical force field developed for the rapid evaluation of mutations' effect on the stability, folding, and dynamics of proteins and nucleic acids (DNA and RNA) and protein-DNA (Blanco et al., 2018, 2019) developed initially by the group of Dr. Luis Serrano. FoldX has been extensively validated in several independent studies, and its force field outperformed Rosetta in 30 out of 33 comparisons made in independent studies.

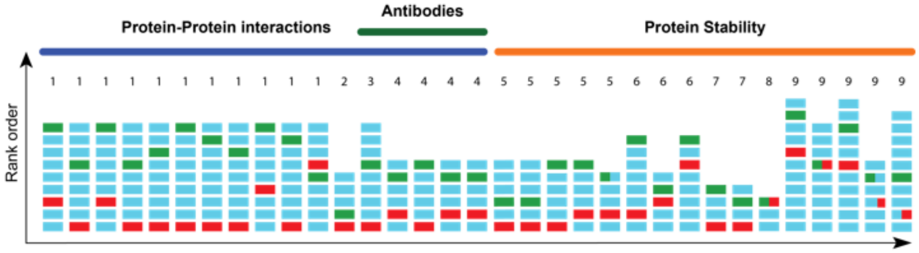


Figure 10. Diagram of publications comparing in silico protein software tools for 33 independent studies. FoldX (green) outperformed Rosetta (red) in 30 comparisons and ties it on two occasions. The figure is taken from a grant of VIB (Vlaams Instituut Voor Biotechnologie) and CRG.

The evolution of FoldX termed ModelX (proprietary of CRG and Dr. Serrano Laboratory) is capable of performing protein backbone engineering and uses a library of protein fragments (>20 million fragments) as blocks that can be used to explore the conformational landscape of a molecule. ModelX now has high-level algorithms that control torsional errors that, combined with FoldX, can be applied to fill structural gaps, ensemble backbones, or transplant protein loops. ModelX is mature enough to be used as a tool to perform protein engineering [redacted]

[redacted]

[redacted]

[redacted]

2. Chapter 2.

Montero-Blay, A., Miravet-Verde, S., Lluch-Senar, M., Piñero-Lambea, C. & Serrano, L. SynMyco transposon: engineering transposon vectors for efficient transformation of minimal genomes. *DNA Res.* **26**, 327–339 (2019).

SynMyco transposon: engineering transposon vectors for efficient transformation of minimal genomes.

Montero-Blay A [1], Miravet-Verde S [1], Lluch-Senar M [1], Piñero-Lambea C [1] [*], Serrano L [1,2,3*]

1 Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003, Spain.

2 Universitat Pompeu Fabra (UPF), Barcelona, Spain.

3 ICREA, Pg. Lluis Companys 23, 08010 Barcelona, Spain.

*To whom correspondence should be addressed. E-mail: luis.serrano@crg.eu, carlos.pinero@crg.eu, Telephone: +34 933160198; +34 933160259

Abstract

Mycoplasmas are important model organisms for Systems and Synthetic Biology, and are pathogenic to a wide variety of species. Despite their relevance, many of the tools established for genome editing in other microorganisms are not available for Mycoplasmas. The Tn4001 transposon is the reference tool to work with these bacteria, but the transformation efficiencies reported for the different species vary substantially. Here, we explore the mechanisms underlying these differences in four Mycoplasma species, *M. agalactiae, M. ferituminatoris, M. gallisepticum* and *M. pneumoniae,* selected for being representative members of each cluster of the Mycoplasma genus. We found that regulatory regions driving the expression of the transposase and the antibiotic resistance marker have a major impact on the transformation efficiencies. We then designed a synthetic regulatory region termed SynMyco RR to control the expression of the key transposon vector elements. Using this synthetic regulatory region, we were able to increase the transformation efficiency for *M. gallisepticum, M. feriruminatoris* and *M. agalactiae* by 30-, 980- and 1036-fold, respectively. Finally, to illustrate the potential of this new transposon, we performed the first essentiality study in *M. agalactiae,* basing our study on more than 199,000 genome insertions

Introduction

The *Mollicutes* class represents a taxonomic group of bacteria that has undergone an extreme genome downsizing process termed degenerative evolution [1]. As a consequence of this evolutionary process, these bacteria are characterized by the lack of a cell wall, streamlined genomes, and very limited biosynthetic pathways. All these features have turned *Mollicutes,* and particularly some species encompassed within the Mycoplasma genus, into appealing models to study basic principles of complex cellular processes such as transcription and translation. For instance, *Mycoplasma pneumoniae* is an important model organism for Systems Biology as highlighted by the comprehensive knowledge acquired about its complete genome [2], transcriptome [3], proteome [4], DNA methylome [5] and gene essentiality [6]. In addition, the first whole-cell computational model developed was for *Mycoplasma genitalium* [7]. At the same time, Mycoplasmas are also interesting organisms for the emerging field of Synthetic Biology and to study the minimal set of genes required to sustain life [6,8–10]. Indeed, the genome of the first synthetic bacterium, JCVI Syn3.0, was inferred from the genome of *Mycoplasma mycoides* following a top-down approach in a design-build and test cycle [11]. Aside from their relevance as model organisms, Mycoplasmas are also a serious concern for the medical and veterinary fields given their pathogenic effects on a wide variety of species. For instance, *M. pneumoniae* is a human pathogen that causes atypical pneumonia[7,12], *Mycoplasma gallisepticum* causes pneumonia in poultry [13], and *Mycoplasma agalactiae* causes contagious agalactia, a common disease in small ruminants with a tremendous economic impact [14].

Identifying genes involved in pathogenicity is critical for generating new vaccines and therapies. Genes that are dispensable for *in vitro* growth (i.e., non-essential, NE) but essential (E) for the infection process could be target genes for vaccine design. Construction of essentiality maps is a fast way to decipher the essential or non-essential character of every single gene found in the genome of a given bacterium. Furthermore, these maps can also help in the identification of those genes that are not essential but whose disruption affects the fitness of the bacteria in a particular environment (i.e., fitness, F). The generation of these maps relies on the construction of saturating libraries of transposon mutants, followed by high-throughput insertion tracking by ultra sequencing (HITS) at different passages [6,15]. Furthermore, construction of saturating libraries is also relevant for Haystack mutagenesis [16], a technique useful to establish gene-function relationships that is widely employed in Mycoplasmas, as a consequence of the paucity of other genome editing tools for these bacteria [17].

Tn4001 is a gentamicin, tobramycin and kanamycin resistance-conferring transposon that was originally found in *Staphylococcus aureus* [18], but has been widely employed in Mycoplasmas [6,10,19]. Few modifications have been made to this transposon to adapt it for use in Mycoplasmas aside from i), placing the transposase coding gene outside the inverted repeats (in plasmids termed mini-transposons) to prevent re-excision from the transposon after the first transposition event [20], and ii) replacing the original gentamicin resistance marker by tetracycline [18], chloramphenicol [21] or puromycin [22] resistance genes. Such a poor adaptation of the Tn4001 transposon to Mycoplasmas might partially explain the dramatic differences observed among transformation efficiencies (TE) in species from this genus. For example, TE is consistently higher in those species belonging to the pneumoniae cluster (i.e., species closely related to *M. pneumoniae*) [23] than in those species encompassed in the spiroplasma [24] or hominis clusters [25].

Here, we explored and identified the reasons underlying the lower transposon TE in a set of different Mycoplasma species that differ in their phylogenetic distance to *M. pneumoniae*. We hypothesized that poor recognition of certain gene regulatory regions in the transposon vector might limit TE in some species. Therefore, we rationally engineered a vector variant that significantly increased the TE in all species tested except for *M. pneumoniae,* which was already transformed at high efficiencies with the unmodified vector. These species were selected to encompass the three different clusters described in the Mycoplasma phylogenetic tree, thereby opening up the door to more global studies of Mycoplasma species. Furthermore, a reporter assay allowed us to identify which of the regulatory regions determining expression of the antibiotic resistance and transposase protein coding genes found in the native transposon vector were inefficiently recognized in each of the strains selected for this work. Finally, to show the potential of our transposon vector, we performed an essentiality study of *M. agalactiae*, obtaining an insertional coverage similar to the one reported for *M. pneumoniae* [6].

Results and discussion

1- Transformation efficiencies in Mycoplasmas show dramatic differences depending on the strain

As a starting point of the project, we aimed to quantify the transposon transformation efficiency (TE) in representative members of each cluster of the Mycoplasma genus (Fig.1A). To this end, cultures of *M. feriruminatoris* (spiroplasma cluster), *M. agalactiae* (hominis cluster), *M. pneumoniae*, and *M. gallisepticum* (pneumoniae cluster) were transformed with the pMTnGm vector [23], a mini-transposon plasmid derived from the original Tn4001 that confers resistance to gentamicin and is broadly employed in the Mycoplasma field. As expected from previous reports, *M. pneumoniae* showed the highest TE, with approximately one transformant for every 10^3 cells. *M. pneumoniae* was followed by *M. gallisepticum* and *M. agalactiae*, with three transformants for every 10^6 cells. Far behind these TEs was that of *M. feriruminatoris*, for which we found one transformant for every 10^9 cells (Fig. 1B).

Since the expression of the gentamicin resistance gene has been found to exert a detrimental effect on growth in *M. genitalium* even in the absence of gentamicin itself [23], we wanted to determine whether this effect was exacerbated in *M. gallisepticum, M. agalactiae and M. feriruminatoris.* This could be responsible for the low TE observed in these species with the pMTnGm vector. To this end, we repeated the same set of transformations, but this time with a modified version of pMTnGm vector termed pMTnCm. In this vector, the native gentamicin resistance gene as well as its regulatory region were replaced by a chloramphenicol resistance under the control of the p438 regulatory region, a minimal 22-bp sequence that controls the expression of a putative restriction enzyme in *M. genitalium.*

For the sake of clarity, from this point on, we will refer to the sequence immediately upstream of the translational start codon of each gene as the regulatory region (RR), comprising the promoter region involved in transcription as well as the 5' UTR involved in translation.

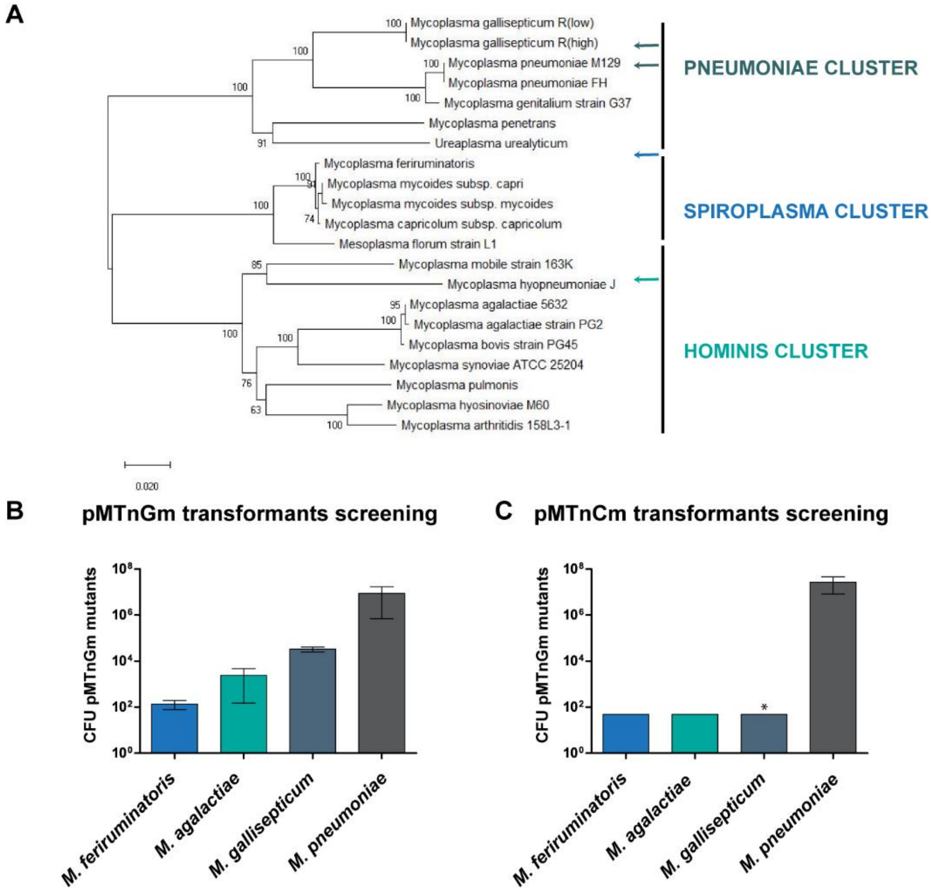


Figure 1. Screening of transposon transformation efficiencies across the mycoplasmal landscape*. A) Phylogenetic tree of 21 selected Mycoplasma species in which three main clusters (pneumoniae, spiroplasma and hominis) can be identified using the Maximum Composite Likelihood method. The tree is drawn to scale with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree* [29]*. B) Bar plot showing the average of gentamicin resistant CFUs (in logarithmic scale) obtained for each of the indicated strains when using the pMTnGm vector (n=3). C) Bar plot showing the average of chloramphenicol resistant CFUs (in logarithmic scale) obtained for each of the indicated strains when using the pMTnCm vector. For the statistical analysis, for those species in which no mutants were detected the number of CFU was set to 49, the maximum value below the limit of detection. One-tailed t-test p-values are indicated with one asterisk (*) when p <0.05 for TE obtained with pMTnGm vector compared to the TE obtained with pMTnCm vector.*

We found that the number of *M. pneumoniae* transformants obtained with the pMTnCm vector was more than three times higher than the number obtained with the pMTnGm vector, thus falling in line with the previous reports of toxicity associated with expression of the gentamicin resistance gene [23]. On the other hand, the TE was dramatically lower in all the other species tested, with the number of total transformants falling under our detection limit of 50 CFU per batch for *M. gallisepticum, M. feriruminatoris and M. agalactiae* (Fig. 1C). The most plausible explanation for these results is that only *M. pneumoniae* is able to efficiently recognize both the regulatory region found upstream of the gentamicin resistance gene and the *M. genitalium*-derived p438 sequence driving the expression of the chloramphenicol resistance gene. In contrast, it seems that *M. gallisepticum, M. feriruminatoris and M. agalactiae* can recognize the regulatory region controlling the gentamicin resistance gene but not the one controlling the chloramphenicol resistance gene.

These results suggest that TE might be directly related to the ability of the transcriptional/translational machinery of each strain to efficiently recognize the regulatory regions controlling not only the antibiotic resistance gene, but also the transposase coding gene.

2- Design of an efficient regulatory region for a broad range of Mycoplasma species.

Several aspects might influence the TE of a transposon vector and can be classified into two different categories. In the first category, we can include those factors that affect the TE of any vector, such as degradation of the plasmid by bacterial restriction machinery, or the optimality of the transformation protocol itself. Leaving these aside, there is a second group of factors that are specific for transposon vectors. Thus, once the vector has entered the cell, proper transcription/translation of the transposase coding gene is required for the correct insertion of the antibiotic resistance gene into the chromosome. Second, the insertion mutant would only be able to grow if the protein levels of the antibiotic resistance gene reach a certain threshold that is sufficient to promote growth on selective medium. Therefore, in an attempt to maximize the protein levels of both the transposase coding gene and the antibiotic resistance gene in all Mycoplasma species, we decided to design a regulatory region termed SynMyco Regulatory Region (SynMyco RR) that would allow efficient transcription and translation in different mycoplasma species.

As a reference for the design, we chose the regulatory region of MPN665, a gene whose transcript is in the top 5% of most transcribed genes in the transcriptome of *M. pneumoniae* [41]. Also, its protein levels (2,646 copies per cell) are among the highest in the of proteome of *M. pneumoniae* [4]. MPN665 encodes for the Elongation factor Thermo-unstable (EF-Tu) protein. As a main component of the ribosome, the protein product of this gene has an essential role in translation and is universally conserved in the prokaryotic world [42,43]. Thus, with the aim of identifying important elements, we aligned the regulatory regions of MPN665 orthologues found in a representative set of Mycoplasma species (Fig. 2A). For the design of the SynMyco RR, we mainly kept those bases found in the MPN665 regulatory region that are not changed to another base in other species. However, given that native regulatory regions did not evolve to be the most productive transcriptional drivers, we also favoured those bases that were found to be optimal as transcription/translation determinants in a recent screen of synthetic sequences [44]. This is exemplified by including an extended Pribnow box motif (TGN-Pribnow) within the SynMyco RR, a feature that is not found in any of the native EF-Tu regulatory regions analyzed but promotes higher transcription rates. Based on the screen mentioned above, we also changed the Pribnow box sequence from TAAAAT to the canonical TATAAT motif. Lastly, to ensure the functionality of the SynMyco RR in a broad range of

Mycoplasma species, we paid special attention to other features found in the regulatory regions such as the -35 box or the Shine-Dalgarno region. In particular, the -35 box seems to have lost its prevalent role as a transcriptional driver during Mycoplasma evolution, as indicated by the absence of any consensus sequence among the most productive regulatory regions found in the screen of synthetic sequences [44]. In fact, only a degenerated -35 box with the sequence TTGANN can be found in the regulatory regions of just 20% of the genes of *M. pneumoniae* [45]. However, as the most prevalent sequence found at the -35 area of the EF-Tu regulatory regions analyzed was TGC, we included this in the SynMyco RR. On the other hand, the Shine-Dalgarno region, which is responsible for ribosome docking onto the mRNA, is present in only 26% of *M. pneumoniae* genes. In contrast, in other species such as *M. agalactiae*, it is found in as much as 73% of the coding sequences (Fig. 2B). For this reason, we also included a Shine-Dalgarno region in the design of the SynMyco RR, using the sequence that is most frequently found in the EF-Tu regulatory regions analyzed (i.e., 5’-AAAGGA-3’). The complete sequence of the SynMyco RR with its main sequence determinants indicated, is shown in Fig. 2C.

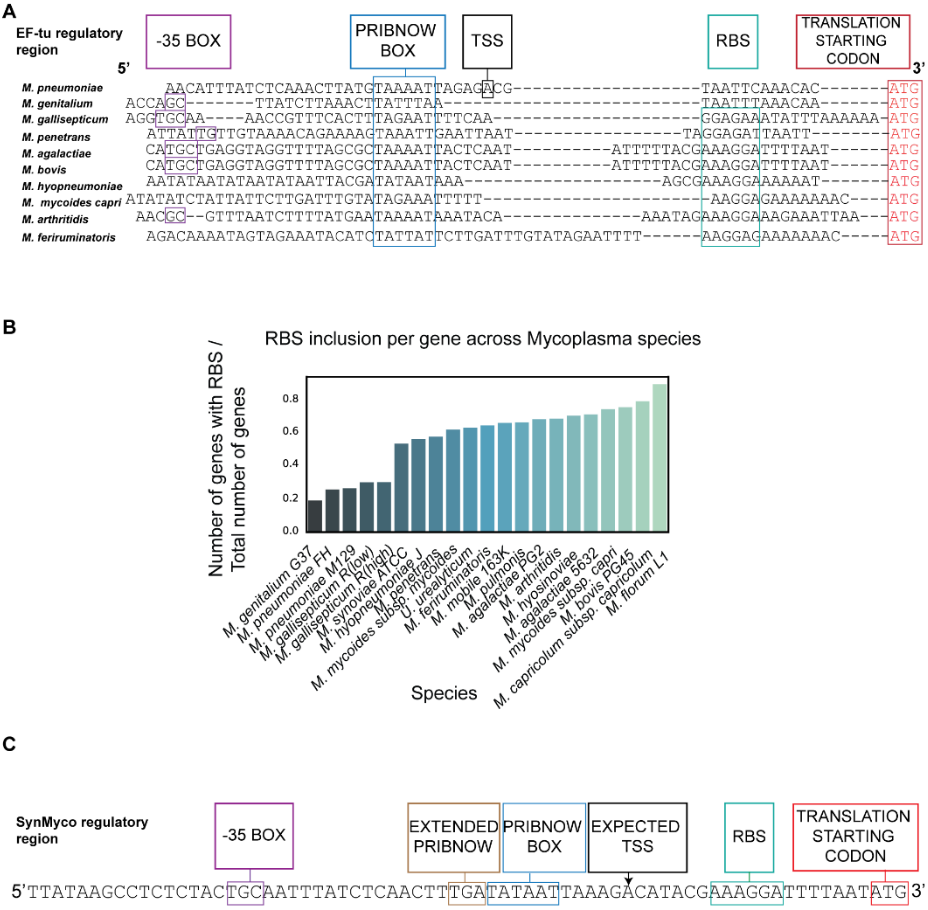


Figure 2. Analysis of regulatory regions found in Mycoplasma species. A) *Sequence alignment of the EF-Tu regulatory regions of ten selected Mycoplasma species. Four main domains are highlighted in boxes: the -35 box, Pribnow box, the RBS sequence and the Translation Starting Codon. In addition, the experimentally determined Transcriptional Start Site (TSS) is also shown for the M. pneumoniae regulatory region.* ***B)*** *Bar plot representing the fraction of RBS-positive genes normalized by the total number of genes per genome in 21 different Mycoplasma species.* ***C)*** *Sequence of the SynMyco regulatory region. The boxes highlight the same domains shown in panel A, plus the extended Pribnow domain, and the expected TSS inferred from the one experimentally determined in M. pneumoniae.*

3- A transposon vector carrying the SynMyco Regulatory Region dramatically increases transformation efficiency

Taking the widely employed pMTnGm vector as reference (Fig. 3A), we constructed a new transposon vector termed pMTnGm-SynMyco in which both the transposase and the gentamicin resistance coding genes were placed under the control of the SynMyco RR (Fig. 3B). Subsequently, we used this vector to determine whether TE is higher with pMTnGm-SynMyco than with pMTnGm. To this end, cultures of *M. pneumoniae*, *M. agalactiae*, *M. gallisepticum,* and *M. feriruminatoris* were transformed in parallel with both vectors.

We did not find significant differences in TE for *M. pneumoniae* when transformed with either of the two vectors (Fig 3C), which is not surprising taking into consideration that *M. pneumoniae* already showed a high TE with the pMTnGm vector. Therefore, these data suggest that the expression levels obtained with the regulatory regions of the pMTnGm vector were already enough to saturate the system in this strain.

On the other hand, the TEs obtained with the pMTnGm-SynMyco vector were significantly higher ($p < 0.05$) in all other species when compared with the TEs obtained with the pMTnGm vector. In particular, for *M. gallisepticum,* we observed a 30-fold increase in TE, obtaining more than 10^6 mutants when transformed with pMTnGm-SynMyco (Fig. 3D).

For *M. feriruminatoris,* the strain showing the worst TE with pMTnGm (less than 100 total transformed cells per replicate), we observed a 980-fold increase in TE when transformed with pMTnGm-SynMyco, resulting in more than 90,000 individual clones carrying a transposon insertion (Fig. 3E).

Lastly, for *M. agalactiae,* we found that whereas with the pMTnGm vector we usually observed slightly more than 200 transformed cells per replicate, we obtained more than $3x10^5$ insertion mutants when transformed with pMTnGm-SynMyco, a value that represents an increase in TE of more than 1036 times (Fig. 3F).

Thus, when transforming these Mycoplasmas strains with the pMTnGm vector, we can classify them into three different groups according to their TEs. The first group is the high efficiency group and contains *M. pneumoniae*, which showed a TE of 10^{-3}. This is followed by the intermediate TE group (i.e. 10^{-5}-10^{-6}) composed of *M. agalactiae* and *M. gallisepticum,* and finally by the low TE group (i.e. 10^{-9}) with *M. feriruminatoris* as the representative. In contrast, when transforming with the pMTnGm-SynMyco vector, we are only able to distinguish a high TE group (composed of *M. pneumoniae*, *M. agalactiae* and *M. gallisepticum*) and an intermediate TE group, (*M.*

feriruminatoris alone).

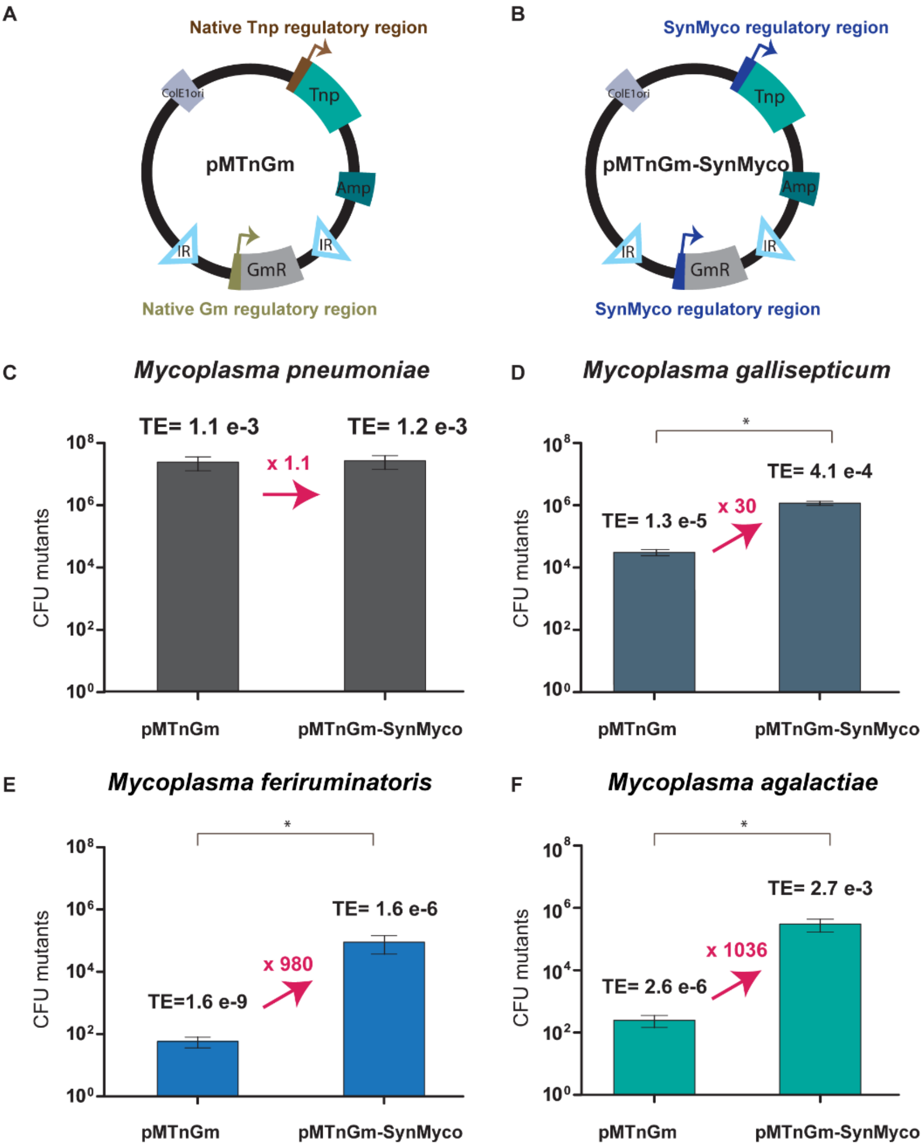


Figure 3. Comparison between the transposon TE obtained with pMTnGm and pMTnGm-SynMyco in four different Mycoplasma species*. A) Scheme of the key modules of the pMTnGm transposon. B) Scheme of the key modules of the pMTnGm-SynMyco transposon. For both A) and B) the abbreviations that appears in the figure are: Tnp for transposase coding gene, Amp for ampicillin resistance coding gene, IR for inverted repeats, ColE1 ori for ColE1 origin of replication and GmR for gentamicin resistance coding gene. C) Bar plot representing the average CFUs of M. pneumoniae resistant to gentamicin (in logarithmic scale) obtained for three independent transformation replicates carried out with either pMTnGm (left side of*

each panel) or pMTnGm-SynMyco (right side of each panel). For each group of bars, the average of TE (CFU resistant to gentamicin / total CFU viable after transformation) is displayed on top. The fold change in TE is indicated over pink arrows connecting both sides of each panel. One-tailed t-test p-values are indicated with one asterisk (*) when p <0.05 for TE obtained with pMTnGm-SynMyco vector compared to the TE obtained with pMTnGm vector. Similar bar plots are shown in D) for M. gallisepticum, E) for M. feriruminatoris and F) for M. agalactiae.

It should be noted that although TE is consistently increased in *M. gallisepticum*, *M. feriruminatoris* and *M. agalactiae* when transformed with pMTnGm-SynMyco vector, the actual numbers of total mutants obtained might be further increased by carefully optimizing the transformation protocol for each species, something that we have not addressed in this study. In addition, while we have generated only a gentamicin version of the SynMyco transposon, we hypothesized that the other resistance markers available for Mycoplasmas (i.e. chloramphenicol, puromycin and tetracycline) [18] and already implemented in native Tn4001 transposon-derived vectors, might in the future be included in pMTnGm-SynMyco vector in substitution of the gentamicin resistance gene. This would allow the generation of a set of four different transposon vectors highly efficient in a broad range of Mycoplasmas. Furthermore, as was recently shown for *M. genitalium*, resistance genes can also be flanked by lox sites in vectors carrying the SynMyco RR thereby allowing the recycling of antibiotic markers when employing protocols that require the iterative use of transposon insertion mutagenesis [46].

4- Strength comparison of SynMyco Regulatory Region and those in the native pMTnGm vector.

In all species tested, the number of transformed cells was higher when transformed with pMTnGm-SynMyco vector than with unmodified pMTnGm vector. These results suggested that the protein yields of the transposase and/or the protein conferring resistance to gentamicin were major determinants of the TE. In order to find out which of the two gene products, or both, were responsible for the increase in TE, we developed a reporter assay. This reporter assay allows quantification of the relative strength of SynMyco RR compared with the regulatory regions driving the expression of the transposase (i.e. Tnp RR) and the gentamicin resistance gene (i.e. Gm RR) in the native pMTnGm vector. To this end, we created three different vectors derived from the pMTnGm-SynMyco in which the genes coding for the mCherry and Venus fluorescent proteins were introduced for quantification and normalization purposes, respectively. The region controlling the expression levels of the mCherry protein varies between the three different vectors, being either SynMyco RR, Tnp RR or Gm RR, whereas the regulatory region controlling the expression of Venus reporter is constant in all the constructs, being always SynMyco RR (Fig. 4A). The three different constructs were transformed in all of the species studied in this work. Subsequently, the ratio between mCherry fluorescence and Venus fluorescence allowed us to determine the relative strength of each regulatory region in *M. pneumoniae* (Fig. 4B), *M. gallisepticum* (Fig. 4C), *M. feriruminatoris* (Fig. 4D) *and M. agalactiae* (Fig. 4E).

In all species, the ratio between mCherry and Venus fluorescence is close to 1 (i.e. 0.9 for *M. pneumoniae*, 1.1 for *M. gallisepticum*, 1.1 for *M. feriruminatoris* and 1.3 for *M. agalactiae*) when both reporters were under control of the SynMyco RR. Interestingly, in *M. gallisepticum* and *M. feriruminatoris,* the ratio is also close to one when mCherry is under control of Gm RR (i.e. 1 for *M. gallisepticum* and 1.1 for *M. feriruminatoris*) but drastically drops when the expression of mCherry is driven by Tnp RR (i.e. 0.2 for *M. gallisepticum* and 0.03 for *M. feriruminatoris).* Altogether, these data suggest that in these two species, when transforming with pMTnGm or pMTnGm-SynMyco vectors, the difference in TE would be due to the expression level of the transposase gene and not the antibiotic resistance gene. Poor performance of Tnp RR was also observed in *M. pneumoniae* and *M. agalactiae* as indicated by the low fluorescence ratio observed when mCherry is under control of Tnp RR (i.e. 0.1 for *M. pneumoniae*, and 0.1 for *M. agalactiae*). Moreover, in contrast to *M. gallisepticum* and *M. feriruminatoris* where Gm RR and SynMyco RR provide similar protein yields, in *M. pneumoniae* and *M. agalactiae* the fluorescence ratio is also reduced when the expression of mCherry is driven by Gm RR (i.e. 0.1 for *M. pneumoniae* and 0.2 for *M. agalactiae*). Thus, in *M. pneumoniae* and *M. agalactiae*, SynMyco RR outperforms not only Tnp RR but also GmRR in terms of protein yields.

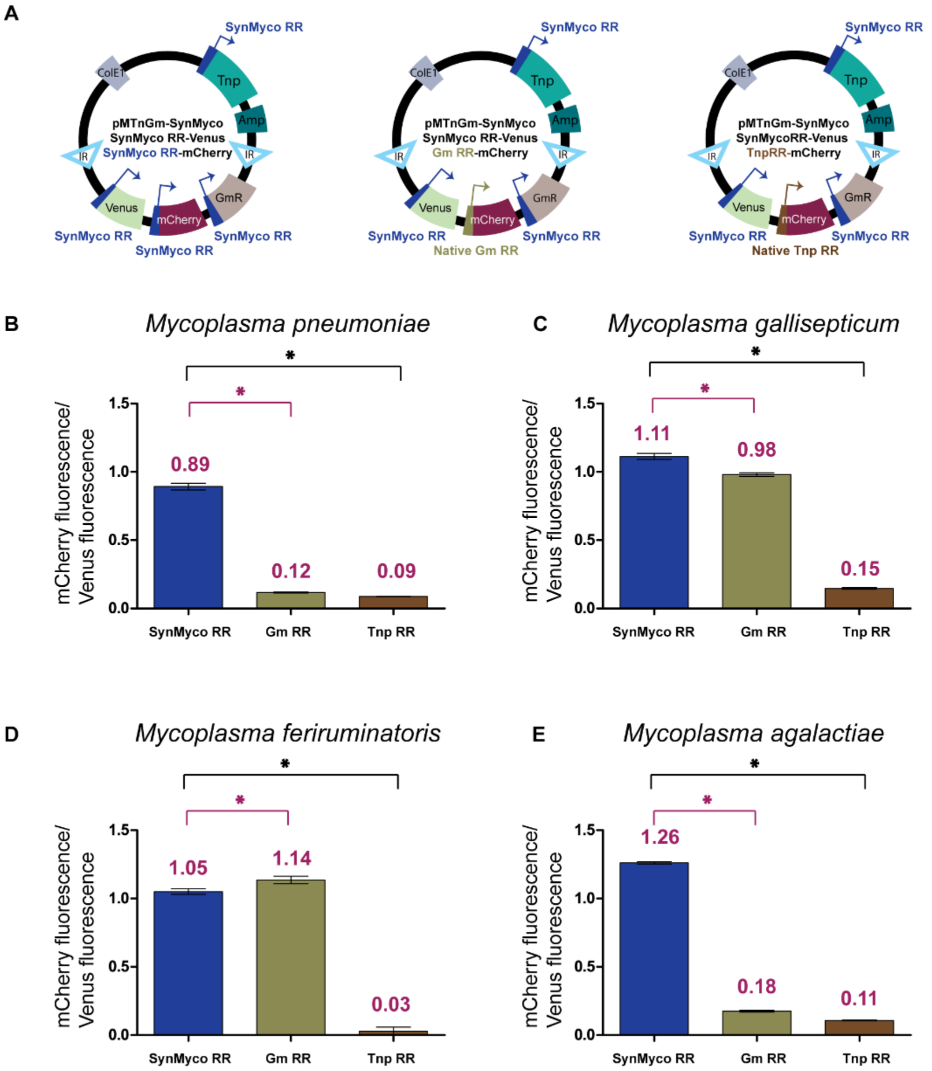


***Figure 4. Comparison of the SynMyco RR efficiency versus the native regulatory regions of pMTnGm transposon**. A) Scheme of the key modules of (i) pMTnGm-SynMyco+SynMyco RR-Venus+SynMycoRR-mCherry transposon, (ii) pMTnGm-SynMyco+SynMycoRR-Venus+Gm RR-mCherry transposon and (iii) pMTnGm-SynMyco+SynMyco-Venus+TnpRR-mCherry transposon. The abbreviations that appear in the figure are: Tnp for transposase coding gene, Amp for ampicillin resistance coding gene, IR for inverted repeats, ColE1 ori for ColE1 origin of replication, GmR for gentamicin resistance coding gene, Venus for Venus protein coding gene and mCherry as mCherry protein coding gene. B) Bar plot representing the ratio obtained in M. pneumoniae for mCherry/Venus fluorescence using transposons represented in panel A. In blue, mCherry coding gene is under the control of SynMyco RR. In gold, mCherry coding gene is under the control of Gm*

RR. In brown, mCherry coding gene is under the control of Tnp RR. On the top of each bar in pink, the average ratio of mCherry/Venus fluorescence obtained for each of the constructs in three replicates. One-tailed t-test p-values are indicated with one asterisk (*) in pink when p <0.05 for the ratio of mCherry/Venus under the control of SynMyco RR versus either Gm RR or Tnp RR. Similar bar plots are shown in C) for M. gallisepticum, D) for M. feriruminatoris and E) for M. agalactiae.

It has not escaped our notice that the reporter assay shows the same expression profile for *M. agalactiae* and *M. pneumoniae*, whereas TE upon using pMTnGm-SynMyco is only drastically increased in *M. agalactiae* (i.e. 1036 fold change) but not in *M. pneumoniae* (i.e. 1.1 fold change). We hypothesized that this observation might be related with the different doubling times of the species. Slow dividing species, such as *M. pneumoniae*, would have more time to deliver the transposon cargo into the chromosome before cell duplication starts to dilute the plasmid within the growing population. In contrast, in species dividing faster, such as *M. agalactiae*, a quick transposition into the chromosome mediated by high expression levels of the transposase coding gene would represent an advantage to avoid the dilution effect associated with cell division and thus would lead to an increased TE. An alternative or complementary hypothesis is that *M. pneumoniae* might be somehow more permissive than the other species for the presence of extrachromosomal elements inside the cell. In this scenario, a fast transposition into the chromosome mediated by high expression levels of the transposase coding gene would represent a greater advantage for the other mycoplasma strains than for *M. pneumoniae.*

In summary, our reporter assay shows that SynMyco RR is efficiently recognized in all the species tested, and suggests that the factor limiting the TE with native pMTnGm vector in most Mycoplasma species is the expression of the transposase coding gene.

5- Unblocking the global study of essential and dispensable genes in *M. agalactiae*

To the best of our knowledge, studies regarding essentiality in Mycoplasmas have only been published so far for *M. genitalium* [8], *M. pneumoniae* [6], *Mycoplasma pulmonis* [47], *Mycoplasma bovis* [10] or *Mesoplasma florum* [48]. However, while for *M. pneumoniae* almost 350,000 unique transposon insertion mutants have been tracked, the other studies analyzed a substantially lower number of clones. For instance, the *M. bovis* study only obtained 319 mutants, representing one insertion every 3,145 bp of the total genome [10], and the study in *M. genitalium* analyzed around 3,300 mutants, showing an insertional coverage of one disruption every 175 bp [8]. Obviously, the accuracy of the assignment of genes to each one of the three categories of essentiality (i.e. essential (E), non-essential (NE), and fitness (F)) is directly related to the insertional coverage of the transposon mutagenesis. Moreover, the existence in bacteria of small open reading frame-encoded polypeptides (i.e., Short open reading frame-Encoded Proteins, also known as SEPs) as short as 11 amino acids in length is becoming widely accepted [49]. Thus, high transposon TE is necessary to study the essentiality of these SEPs [6,50].

To illustrate the potential that our pMTnGm-SynMyco vector could have in the study of a broad range of Mycoplasma species, we decided to perform an essentiality study for *M. agalactiae* 7784, a strain for which no essentiality map is currently available. First, we sequenced, assembled and annotated for the first time the genome of this strain. Then, after transforming the strain with pMTnGm-SynMyco vector we were able to map 199,723 transposon insertions to the *M. agalactiae* genome, with a estimated genome size of 853,960. This represents ~23.3 insertions every 100 bp, which is around half of the coverage observed in *M. pneumoniae* [6] (354,447 transposon insertions mapped, representing ~ 43 insertions every 100 bp) but still substantially higher than the ones reported in other essentiality studies (Fig. 5A and Fig. 5B).

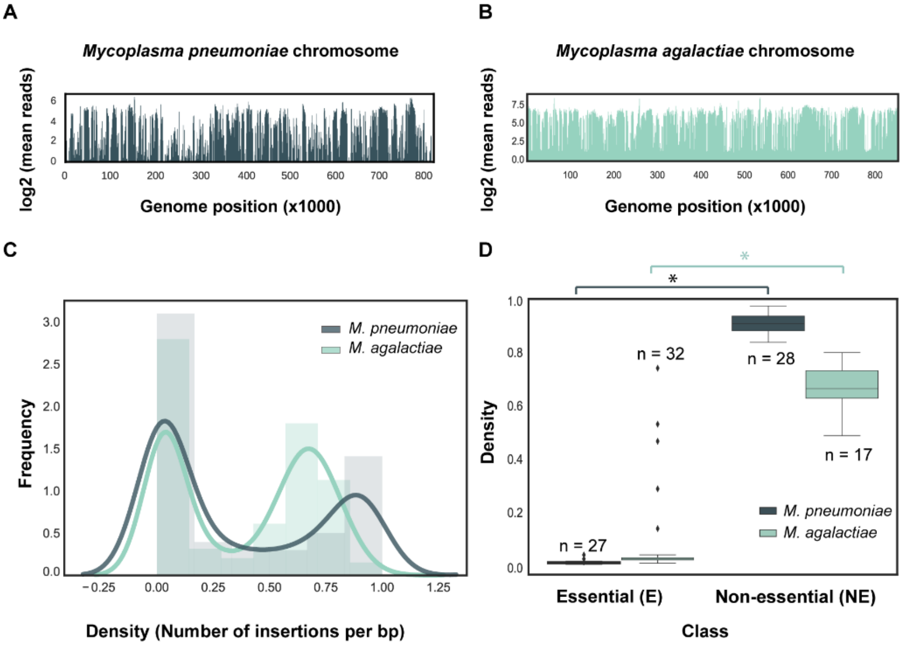


Figure 5. Essentiality study in M. agalactiae using the pMTnGm-SynMyco transposon and a comparison with previous studies in M. pneumoniae. A) Genome disruption profile for M. pneumoniae. The y-axis represents the logarithmic average of total reads covering a window of 1,000 bp (x-axis). B) Genome disruption profile for M. agalactiae representing the same information as in the previous panel. C) Insertion density by gene distribution in M. pneumoniae (dark blue) and M. agalactiae (light blue). The x-axis represents the percentage of bp in a gene that are disrupted and the y-axis the frequency of densities in the distribution. To better compare M. pneumoniae and M. agalactiae transposon insertion distributions, we standardized both distributions using min-max scaling. D) Box-plot representing the statistical comparison of specific subsets of genes expected to be essential (E) and non-essential (NE) in M. pneumoniae (gray) and M. agalactiae (green). The asterisk in gray represents p-value < 0.05 (3.62 e^{-41}) when comparing density of insertions of E and NE coding genes in M. pneumoniae. The asterisk in green represents p-value < 0.05 (1.20 e^{-20}) when comparing density of insertions of E and NE coding genes in M. agalactiae.

When considering the frequency of insertions per gene, we observed a bimodal distribution separating essential from non-essential genes in both strains (Fig. 5C). Next, we explored the insertional profile at the gene level using sets of known E and NE genes as inferred from the reference essentiality study in *M. pneumoniae* [6]. Whereas for *M. pneumoniae* the training sets for E and NE genes comprise 27 and 28 genes respectively, for *M. agalactiae* we generated dedicated training sets based on bibliography containing 32 E genes and 17 NE genes. When comparing the average density of insertions between E and NE genes in *M. pneumoniae* and *M. agalactiae,* we observed that the two groups were significantly different within each species (two-tailed t-test with equal variances, p-values equal to $1.42e^{-47}$ and $1.05e^{-20}$, respectively; Fig. 5D). Taken altogether, we were able to assign one category for 689 genes found in the *M. agalactiae 7784* genome: E (43.98% of genes), F (25.25% of genes) or NE (30.77% of genes) (see supplementary table 8), in line with the percentage of E (49.28%), F (13.40%) and NE (37.32%) genes obtained for *M. pneumoniae* [6].

When *M. agalactiae* essential genes were classified according to the cluster of orthologous groups (COG) functional categories annotation system, we found that the categories that were significantly enriched in essential genes were: (i) protein coding genes involved in translation, ribosomal structure and biogenesis, (ii) functional RNAs, (iii) protein coding genes without assigned COG category and (iv) protein coding genes with unknown function (see Supplementary table 8).

However, for the purposes of this work it is more relevant to compare the essentiality map of *M. agalactiae* at high density of insertions with one obtained with low density to illustrate the importance of high coverage. Unfortunately, there is no essentiality study done in *M. agalactiae,* but there is a low coverage analysis in its closely related specie *M. bovis* PG45 [10] (319 transposon insertional mutants individually sequenced, versus more than 199,000 transposon insertions analyzed by deep sequencing in our study). Sequencing of individual insertion clones is useful to classify the genes within the F-NE categories (i.e. genes disrupted by a transposon insertion) but cannot distinguish between both. Examples of this limitation are the genes coding for a tRNA modification GTPase and for the deoxyribonuclease IV. In the *M. bovis* essentiality study, both genes (i.e. MBOVPG45_0060 and MBOVPG45_0301) were classified as NE given the isolation of individual clones carrying insertions within their coding region. However, our essentiality study based on the analysis of at least 600 times more mutants than the one of *M. bovis* and mapped by ultra sequencing classified these two genes (i.e. MAGA7784_RS00280 and MAGA7784_RS02715) as F in *M. agalactiae*. This implies that although these genes are non-essential, their disruption cause a growth impairment of these particular clones at least for *M. agalactiae*, something that cannot be directly measured by genomic sequencing of transposon insertion in a limited number of clones.

Moreover, aside from its lack of accuracy in distinguishing between NE and F genes, the main limitation of small transposon libraries is related with E genes. Specifically, the low coverage of these libraries makes it difficult to ascertain whether a gene is free of insertions because of its essential character or as a result of the randomness of the integration and/or low sampling of the mutants. Indeed, only two genes were suggested by the authors as highly probable essential genes in the *M. bovis* study, MBOVPG45_0337 and MBOVPG45_0710, coding for an ATP-binding protein and SGNH/GDSL hydrolase, respectively. This assumption was based on the lack of clones carrying a transposon insertion within the coding regions of these genes in spite of their large size (i.e. 3420 bp for MBOVPG45_0337 and 8013 bp for MBOVPG45_0710). Nonetheless, it should be noted that the insertional coverage of the mutant library of this study, with one insertion every 3145 bp of the genome on average, is in the same range of these genes in terms of size. In our *M. agalactiae* essentiality study, the orthologue of the SGNH/GDSL hydrolase coding gene (i.e. MAGA7784_RS03410) was found E, confirming the hypothesis of the *M. bovis* report. In contrast, the gene coding for the ATP binding protein (i.e. MAGA7784_RS02615) was determined as NE as indicated by the presence of 891 transposon insertions within its sequence. This suggests that the absence of clones carrying an insertion within MBOVPG45_0337 gene is most likely a consequence of the low coverage of the insertional library, rather than the gene being truly essential in *M. bovis.* Furthermore, the limitation of low coverage libraries to assign a truly E character for a given gene is even more evident with small coding sequences. For instance, in *M. bovis* there were no clones found to carry a transposon insertion within the coding sequences of genes MBOVPG45_0043 (582 bp) or MBOVPG45_0596 (963 bp), coding for sigma-70 RNA polymerase factor and Holliday junction branch migration DNA helicase ruvB, respectively. This might indicate that these genes are E. However, homologues of these genes are fully covered with transposon insertions not only in *M. agalactiae*, (MAGA7784_RS00215 and MAGA7784_RS01205) but also in *M. pneumoniae* (MPN626 and MPN536), suggesting the lack of insertions in the *M. bovis* study is more likely related with the small size of the genes rather than with their essential character. All of these examples illustrate the importance of high coverage transposon insertion libraries for the appropriate category assignment of a given gene, something that can be only achieved with efficient transposon vectors such as pMTnGm-SynMyco.

In conclusion, we demonstrate that the regulatory regions driving the expression of the transposase and the resistance genes have a tremendous impact on the TE achieved after vector transformation. Although problems derived from poor recognition of vector regulatory regions can be avoided with transformation procedures based on purified transposases [51], it should be noted that not all transposases are commercially available and their cost might limit use in many research groups. Indeed, screening libraries of regulatory regions for the key elements of transposon vectors has already been shown to be a useful strategy to produce a moderate increase (i.e., around one order of magnitude) in the TE of different bacteria [52]. However, an approach such as the one followed in our study involving rational design of regulatory regions rather than the screen of a limited number of variants might be more effective in increasing TE. In fact, we have seen that the pMTnGm-SynMyco vector significantly increases the TE in three phylogenetically distant Mycoplasma species. For this reason, we hypothesize that this vector might improve the reported TE in all species belonging to the Mycoplasma genus. Even though we have shown that the SynMyco regulatory region is efficiently recognized in different species, it cannot be excluded that other designs of regulatory regions could also lead to an increase in TE. As exemplified in this work with *M. agalactiae*, the higher TE obtained using this vector may unblock the development of new essentiality maps for other Mycoplasma species, and thereby promote global knowledge of these interesting microorganisms. In addition, the higher insertional coverage obtained with the pMTnGm-SynMyco vector should facilitate the isolation of clones of interest from libraries of insertional mutants, an advancement which may boost gene-function assignments in Mycoplasma species that have been poorly studied so far.

Materials and methods

Bacteria strains and culture conditions.

For wild-type *M. pneumoniae,* the strain M129 (ATTC 29342, subtype 1, broth passage no. 35) was used. Wild-type *M. agalactiae* 7784 was kindly provided by Christine Citti. Wild-type *M. gallisepticum* R high was kindly provided by Michael Szostak [25]. Wild-type *Mycoplasma feriruminatoris* G5847 was kindly provided by Carole Lartigue [26]. All strains were grown at 37ºC in standard Hayflick medium supplemented with 100 µg/mL ampicillin and phenol red (0.005% w/v). *M. agalactiae* standard Hayflick culture was supplemented with sodium pyruvate (0.5%, pH 7.6, Sigma-Aldrich). *M. feriruminatoris, M. agalactiae and M. gallisepticum* were grown in suspension (180 rpm, 37ºC). *M. pneumoniae* was grown without shaking at 37ºC and with 5% CO_2.

Plasmids.

The plasmids were generated following the Gibson assembly method [27]. The list of plasmids as well as a detailed description of the procedure followed to build all the constructs are described in Supplementary table 1. For plasmid generation, DNA was isolated from NEB® 5-alpha High Efficiency (C2987P). The clones were isolated using LB agar + ampicillin (100 µg/mL) plates and confirmed by sequencing (GATC biotech). The list of all the primers used in this study for the generation of the plasmids can be found in Supplementary Table 2.

Transformation protocol.

Transformation procedures were initially based on methods previously described but later slightly modified [24]. For *M. feriruminatoris*, *M. agalactiae* and *M. gallisepticum* cultures, 10-ml log-phase cultures were harvested at 10,000 g for 10 minutes at 4ºC. The medium was removed and 10 ml of fresh Hayflick was added. After three hours, the culture was centrifuged at 10,000g at 4ºC and then washed three times with chilled electroporation buffer (EB; 272 mM sucrose, 8 mM HEPES, pH 7.4) before final resuspension in 300 µl chilled EB. After mixing 50 µL of cells with 1.5 µg of DNA and incubating for 20 minutes on ice, the mix was transferred into 0.1-cm electro cuvettes and electroporated in a BIO-RAD Gene Pulser Xcell apparatus. For *M. pneumoniae* (i.e. adherent strain) cells were grown in a 75-cm^2 tissue flask containing 20 mL of fresh Hayflick and incubated at 37ºC under 5% CO_2 until late exponential phase. Cells were washed twice, resuspended in precooled EB, scraped off and passed through a 25-gauge (G25) syringe needle 10 times. Aliquots of 50 µl of cells in 0.1-cm cuvettes with 1.5 µg of the corresponding plasmid were kept on ice during 20 minutes. The electroporation settings were common for all strains: 1250 V / 25 µF / 100 Ω. Immediately after the pulse, 420 µl of fresh Hayflick was added to the cells. Subsequently, the cells were incubated at 37ºC before seeding on agar plates. The incubation time was 30 minutes for *M. feriruminatoris*, 90 minutes for *M. agalactiae* and *M. gallisepticum* and 120 minutes for *M. pneumoniae*.

Determination of transformation efficiency in Mycoplasma species.

After incubating the transformations at 37ºC during the above-mentioned time depending on the strain, 10-fold serial dilutions of the cultures were performed (from -1 to -8). The dilutions were made in a total volume of 100 µl and 10 µl of each dilution was plated onto Hayflick 0.8% agar plates. Transformations done with pMTnGm or pMTnGm-SynMyco vectors were counted on agar plates supplemented with 100 µg/mL gentamicin. Transformations done with pMTnCm vector were selected in agar plates supplemented with 20 µg/mL chloramphenicol. The mutant counts refer to mutants per transformation, where the final volume is 500 µl. Transformation efficiency (TE) is defined as the ratio of colony forming units (CFUs) counted on antibiotic supplemented plates (i.e., transformed cells carrying a transposon insertion) to non-supplemented plates (i.e., viable cells after transformation). To assess differences in TE between pMTnGm and pMTnCm, one-tailed paired t-tests were applied to three different experimental replicates for each strain employed in the study. As the mutants obtained for *M. agalactiae*, *M. gallisepticum* and *M. feriruminatoris* using pMTnCm transposon were below the limit of detection, for the statistical analysis the number of mutants obtained was set to 49 CFU per batch, a value that corresponds to the maximum number of CFU under the limit of detection for each species. The p-values obtained for TE differences between pMTnGm and pMTnCm were 0.470, 0.186, 0.026 and 0.067 for *M. feriruminatoris*, *M. agalactiae*, *M. gallisepticum* and *M. pneumoniae*, respectively (for the raw data of these transformations as well as the statistical analysis see supplementary table 3). The same statistical approach was applied to compare TE obtained with pMTnGm versus pMTnGm-SynMyco in three different experimental replicates for each species employed in the study (p-values of $1.49 x 10^{-4}$, $4.28 x 10^{-2}$, $4.97 x 10^{-2}$, and 0.325 for *M. feriruminatoris*, *M. agalactiae, M. gallisepticum* and *M. pneumoniae* respectively). Data of TEs obtained with pMTnGm and pMTnGm-SynMyco as well as the statistical analysis are shown in supplementary table 4.

Phylogenetic analysis of selected Mycoplasma species

For the generation of the phylogenetic tree, 21 Mycoplasma species were selected. DNA sequences encoding the 16S rRNA of each species were aligned using multiple sequence alignment by ClustalW. For those species containing more than a single copy coding for the 16S rRNA we selected one of the copies arbitrarily. The evolutionary history of these microorganisms was inferred using the Neighbor-Joining method [28]. The optimal tree with the sum of branch length = 0.91 is shown. The percentages of replicate trees in which the associated taxa clustered together in the bootstrap test (500 iterations) are shown next to the branches. The evolutionary distances were computed using the Maximum Composite Likelihood method [29,30] and are in the units of the number of base substitutions per site. Codon positions included were 1st + 2nd + 3rd + Non coding. All positions containing gaps and missing data were eliminated. There were a total of 1,464 positions in the final dataset. Evolutionary analyses were conducted in MEGA X [31]. The 21 different DNA sequences coding for 16S rRNA and the accession number of the species to which they belong are listed in Supplementary Table 5, except for *M. feriruminatoris* whose assembled genome sequence was kindly provided by Dr. Carole Lartigue.

Ribosome Binding Site inclusion along different Mycoplasma species.

As a reference we used the same set of 21 Mycoplasma species that were used to build up the phylogenetic tree. Specifically, for each species we extracted the 15 bases before the start codon of all their annotated genes. Within these sequences, we then checked for the presence of any of the subsequences reported to be RBS [32]. This list included: GGA, GAG, AGG, AGGA, GGAG, GAGG, AGGAG, GGAGG, AGAAGG, AGCAGG, AGGAGG, AGTAGG, AGGCGG, AGGGGG, and AGGTGG. Inclusion was represented as the percentage of genes in a bacterial species that included one of these RBS motifs (for the percentage of RBS inclusion of the species of the work see Supplementary table 6).

Strength evaluation of native regulatory regions driving the expression of the transposase and gentamicin resistance gene in pMTnGm, and comparison with SynMyco regulatory region in different Mycoplasma species.

Three different reporter plasmids containing the mCherry coding sequence under the control of the three different regulatory regions (i.e. pMTnGm-SynMyco+SynMyco RR-Venus+SynMyco RR-mCherry, pMTnGm-SynMyco+SynMyco RR -Venus+ GmR-mCherry and pMTnGm-SynMyco+ SynMyco RR-Venus+Tnp RR-mCherry) were generated. These plasmids contain genes coding for two fluorescent proteins (i.e. Venus and mCherry). These proteins have fused in its N-terminal part the mp-200 sequence, a 29 amino acid signal from *M. pneumoniae* MPN391a gene that has been previously fused to Venus and mCherry sequences to improve protein stability [33]. Whereas the gene coding for the Venus fluorescent protein was included for normalization purposes in all constructs under the control of SynMyco RR, the gene coding for mCherry has been placed under the control of three different regulatory regions depending on the construct: a) the SynMyco RR, b) the 150 bp upstream region of the transposase coding gene (Tnp RR) and c) the 150 upstream region of the gentamicin resistance coding gene (Gm RR).

All constructs were transformed in *M. agalactiae, M. gallisepticum, M. feriruminatoris* and *M. pneumoniae* following the protocol described above. To generate a primary stock of the transformations, from each 500 µl batch 100 µl was inoculated in a flask containing 5 mL of Hayflick medium supplemented with 100 µg/mL gentamicin for *M. pneumoniae,* while for the other non-adherent strains the 100 µl was inoculated in 50 ml Falcon tubes (Fisher Scientific, 14-432-22) filled with 10 mL of Hayflick medium supplemented with 100 µg/mL gentamicin. When cultures were grown, the total biomass was resuspended in one mL of Hayflick medium and stored at -80ºC.

To determine the fluorescence levels of each construct, cultures of the four different Mycoplasma strains carrying the three different reporter constructs plus a negative control for each strain not carrying any construct were used. The cultures were grown as described above using 10 µl of their respective primary stocks as inoculum (i.e. around 12 h for *M. feriruminatoris,* 48 h for *M. gallisepticum* and *M. agalactiae* and 72 h for *M. pneumoniae*) and cells were harvested and washed twice with chilled PBS buffer until final resuspension of the cultures in 500 µl of PBS. Five-fold serial dilutions of the final cell suspensions were done, and 100 µl of all dilutions were loaded in 96 Well Optical Btm Plt Polymerbase Black Lid plates (165305, Thermo Scientific).

The absorbance and fluorescence values were measured using Tecan I-control 1.9.17.0 Infinite 200. The settings were determined for optimal gain, 25 flashes and 20 µs of integration time. The fluorescence settings were ƛex = 514 nm and ƛem = 574 nm for Venus and ƛex = 550 nm and ƛem = 630 nm for mCherry. For each strain, the absorbance at ƛ= 600 nm was measured. For the fluorescence analysis we took the data of those wells (i.e. dilutions) in which OD_{600nm} absorbance values were between 0.075-0.2; 0.2-0.35; 0.075-0.4; and 0.22-0.66 for *M. feriruminatoris, M. agalactiae, M. pneumoniae* and *M. gallisepticum* respectively. These dilutions were selected so that fluorescent signals were clearly different from negative controls of each strain (i.e. not carrying fluorescent constructs) and proportional to the absorbance (i.e. not saturating signals). Fluorescence arbitrary units (AU) measured for Venus and mCherry were normalized to OD_{600nm} for each condition (Venus AU / OD_{600} and mCherry AU / OD_{600}) to obtain normalized fluorescence AU. For each of the four strains of the work, the mCherry and Venus fluorescence levels of WT cells (i.e. not carrying any fluorescent construct) were determined and subtracted from the fluorescence values obtained for each condition, to obtain subtracted AU. Finally, to compare the strength of all regulatory regions analyzed we calculated for each strain the ratio between normalized and subtracted mCherry UA / normalized and subtracted Venus UA. Statistical paired t-test analysis with one-tailed distribution was performed for each strain comparing the ratio of mCherry AU/ Venus AU for a) SynMyco-mCherry vs Gm RR-mCherry and b) SynMyco-mCherry vs Tnp RR-mCherry. For the data of the strength evaluation of SynMyco RR and the native RR driving the expression of the gentamicin and the transposase coding genes in the native pMTnGm as well as the statistical analysis see Supplementary table 7.

M. agalactiae pMTnGm-SynMyco transformation for essentiality study.

M. agalactiae was transformed using the pMTnGm-SynMyco transposon following the protocol described above. Two hours after the transformation,

⅘ parts of the total 500 µl batch were inoculated into 10 mL Hayflick + 0.5% sodium pyruvate (Sigma-Aldrich P8574-5G) + 100 µg/mL gentamicin. After 24 hours at 37ºC, the culture was centrifuged for 10 minutes at 10,000g. The supernatant was discarded and the cells were collected in 500 µl of Hayflick constituting passage 1 of the transformation. The procedure described above was repeated two more times until passage 3, using a volume of 8 µl of the immediately preceding passage, to grow all the passages (1/62.5 dilution). Taking into account the doubling time of *M. agalactiae* (4-5 hours), and the passages performed the expected number of cell divisions in passage 3 is 20. From passage 3, 350 µl out of the total 500 µl were taken to extract genomic DNA using the MasterPure Complete DNA & RNA Purification Kit (MC85200, Lucigen) following the protocol described by the manufacturer.

Genome assembly and de novo annotation for *M. agalactiae* 7784.

M. agalactiae was grown 48 hours in 25 mL of Hayflick medium supplemented with sodium pyruvate at 37°C as previously described. After, the genomic DNA was isolated using the MasterPure Complete DNA & RNA Purification Kit (MC85200, Lucigen) following the protocol described by the manufacturer. The genomic DNA was sheared to 200-300 bp fragments using a Covaris S2 device. Then, paired-end Illumina libraries were created following previously described protocols [34] and the size selected was 125 bp. The resulting libraries were quantified on an Agilent Bioanalyzer chip (Agilent Technologies). Double-stranded templates were amplified and sequenced on an Illumina GAII. Raw reads were analyzed using the FastQC tool (website: http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc) for assessing the quality and the presence of adapters. Contigs were *de novo* assembled using ABySS [35] and *M. agalactiae* 5632 genome as annotation reference resulting in 159 contigs with average length of 5,511 base pairs. In total, 503 genes present in *M. agalactiae* 5632 displayed homologs within contigs of *M. agalactiae* 7784. We increased this value up to 689 with a specific BlastN search where we restrictively selected hits with an alignment length greater than 95% and a e-value less than $1x10^{-5}$. This allowed us to provide a more accurate list of genes presented by the strain of interest, to assign putative functions by homology as well as to detect gene duplications in the studied genome. We observed that 12 contigs were informative enough to capture a set of 689 genes. This set of contigs averaged 71,163 bp each in length and 853,960 bp in total. The latter value was considered to be the genome size for *M. agalactiae* 7784. The raw data of DNAseq, genome assembly and *de novo* annotation has been submitted as BioProject under accession PRJNA528179.

Sequencing of *M. agalactiae* 7784 transformed with pMTnGm-SynMyco.

Genomic DNA sequencing was performed in the Genomics facility at Centre for Genomic Regulation in a HiSeq Sequencing v4 Chemistry controlled by Software HiSeq Control Software 2.2.58. Settings, 125 nucleotides in paired-end format. In the HiSeq sequencing technology from Illumina Genome Analyzer, the protocol starts with DNA fragmentation. Then, the fragmented DNA is amplified using oligos that add adapters allowing the subsequent binding of PCR products to the glass flow cell. Later, the sequencing is performed by synthesis cycles, in which a single complementary base for each deoxynucleotide (dNTP) is incorporated using a fluorescently labeled dNTP. Finally, lasers excite the fluorophores while a camera captures images of the flow cell.

Transposon mapping.

For *M. agalactiae* 7784, paired-end sequencing raw reads were filtered to remove PCR duplicates using Fastuniq [36]. Then, a specific Inverted Repeat (IR) associated to the transposon insertion process (TTTTACACAATTATACGGACTTTATC, length=26) was trimmed by Trimmomatic [37] and then mapped to the reference using Bowtie2 allowing 1 mismatch [38]. Later, we selected paired reads mapped unambiguously with a minimum alignment quality of 30 using SAMtools [39]. The last step relied on shell text processing tools (awk/grep/sed) to identify those pairs where one of the reads presented a shorter length than the original read length minus 26 (expected length if the IR was removed). Position reported represents the first mapped position in the chromosome contiguous to the IR.

Essentiality definition and training set generation for *M. agalactiae* and *M. pneumoniae*

The essentiality studies were developed using as reference the study previously done in *M. pneumoniae* [6]. Specifically, we reanalyzed the T4 dataset of *M. pneumoniae* [6]. Transposition events are considered to behave as random events resembling a Poisson process over large portions of the chromosome with a uniform density. The analysis starts with two different linear insertion densities (r) for essential (r_E) and non-essential genes (r_{NE}). These values correspond to the total number of insertions mapped normalized by gene length for a training set of genes which, based on prior knowledge, we assume are either E or NE and can be used as reference to classify the rest of genes. For *M. pneumoniae,* we used as a training set the same group of genes that were defined in its essentiality study [6]. In the case of *M. agalactiae*, a new E training set containing 32 genes was generated using the same criteria as for *M. pneumoniae*. For NE genes, we extracted 84 genes with no homolog in four closely related species to *M. agalactiae: Mycoplasma hyosynoviae*, *Mycoplasma arthritidis, Mycoplasma pulmonis,* and *Mycoplasma synoviae*. Out of this group, we selected 17 genes that had been confirmed as NE in the essentiality study done in *M. bovis* PG45 [10] . The probability of a specific gene to be essential or not is evaluated predicting two values: P_E and P_{NE} using formula (1). In this formula, *N* represents the number of insertions mapped to a gene and L the length of that gene. L corresponds to the inner 90% part of each gene (removing the 5% in each terminus) since it has been observed that these regions allow a higher number of non-disruptive mutations as they would not be located in the core regions of the encoded protein. The value of *r*, inferred from the training sets, varies between essential (r_E) and non-essential genes (r_{NE}) and allows us to determine the probability of a particular gene to be essential or not with its specific N and L. If $P_E > 0.0$ and $P_{NE} = 0.0$ the gene is classified as E (essential), $P_E = 0.0$ and $P_{NE} > 0.0$ will correspond to NE (non-essential), and if both probabilities are non-zero we assume that the gene is F (fitness).

$$P_N(L) = \frac{(rL)^N}{N!}e^{-rL} \qquad (1)$$

After, Clusters of Orthologous Groups (COGs) were associated to the *M. agalactiae* genome using eggNOG-mapper [40] utilizing the DIAMOND mapping mode, taxonomic scope set on Firmicutes and prioritizing quality over coverage. A subsequent step of manual curation of COG categories was performed. For the set of 689 genes found for *M. agalactiae* 7784, the genes included in the training list, the COG category assignment, their essentiality assigned class and the statistical analysis see supplementary table 8.

Data availability

The raw data of DNAseq, genome assembly and *de novo* annotation was submitted as BioProject under accession PRJNA528179. The transposon sequencing in *Mycoplasma agalactiae* dataset can be found in ArrayExpress with accession number: E-MTAB-7425.

Authors’ contributions

LS, CPL and AMB conceived the project and the experimental design. AMB performed the experiments. SMV performed the data analysis. LS, MLS and CPL contributed with ideas and direct supervision of the project. All authors participated in evaluating results and discussions about the project. AMB, CPL and SMV wrote the paper, prepared tables and figures. LS and MLS reviewed the manuscript. All authors read and approved the final manuscript.

Bibliography

1. Woese, C. R., Maniloff, J., and Zablen, L. B. 1980, Phylogenetic analysis of the mycoplasmas. *Proc. Natl. Acad. Sci. U. S. A.*, **77**, 494–8.
2. Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B. C., and Herrmann, R. 1996, Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res.*, **24**, 4420–49.
3. Guell, M., van Noort, V., Yus, E., et al. 2009, Transcriptome complexity in a genome-reduced bacterium. *Science*, **326**, 1268–71.
4. Maier, T., Schmidt, A., Guell, M., et al. 2014, Quantification of mRNA and protein and integration with protein turnover in a bacterium. *Mol. Syst. Biol.*, **7**, 511–511.
5. Lluch-Senar, M., Luong, K., Lloréns-Rico, V., et al. 2013, Comprehensive methylome characterization of *Mycoplasma genitalium* and *Mycoplasma pneumoniae* at single-base resolution. *PLoS Genet.*, **9**, e1003191.
6. Lluch-Senar, M., Delgado, J., Chen, W.-H., et al. 2015, Defining a minimal cell: essentiality of small ORFs and ncRNAs in a genome-reduced bacterium. *Mol. Syst. Biol.*, **11**, 780.
7. Karr, J. R., Sanghvi, J. C., Macklin, D. N., et al. 2012, A whole-cell computational model predicts phenotype from genotype. *Cell*, **150**, 389–401.
8. Glass, J. I., Assad-Garcia, N., Alperovich, N., et al. 2006, Essential genes of a minimal bacterium. *Proceedings of the National Academy of Sciences*, **103**, 425–30.
9. French, C. T., Lao, P., Loraine, A. E., Matthews, B. T., Yu, H., and Dybvig, K. 2008, Large-scale transposon mutagenesis of *Mycoplasma pulmonis. Mol. Microbiol.*, **69**, 67–76.
10. Sharma, S., Markham, P. F., and Browning, G. F. 2014, Genes found essential in other mycoplasmas are dispensable in *Mycoplasma bovis*. *PLoS One*, **9**, e97100.
11. Hutchison, C. A., 3rd, Chuang, R.-Y., Noskov, V. N., et al. 2016, Design and synthesis of a minimal bacterial genome. *Science*, **351**, aad 6253.
12. Waites, K. B., and Talkington, D. F. 2004, *Mycoplasma pneumoniae* and its role as a human pathogen. *Clin. Microbiol. Rev.*, **17**, 697–728.
13. Levisohn, S., and Kleven, S. H. 2000, Avian mycoplasmosis (*Mycoplasma gallisepticum*). *Rev. Sci. Tech.*, **19**, 425–42.
14. Kumar, A., Rahal, A., Chakraborty, S., Verma, A. K., and Dhama, K. 2014, *Mycoplasma agalactiae*, an etiological agent of contagious agalactia in small ruminants: A Review. *Vet. Med. Int.*, **2014**, 286752.
15. De Jesus, M. A., Gerrick, E. R., Xu, W., et al. 2017, Comprehensive essentiality analysis of the *Mycobacterium tuberculosis* genome via saturating transposon mutagenesis. *MBio*, **8**.
16. Halbedel, S., Busse, J., Schmidl, S. R., and Stülke, J. 2006, Regulatory protein phosphorylation in *Mycoplasma pneumoniae*. A PP2C-type phosphatase serves to dephosphorylate HPr (Ser-P). *J. Biol. Chem.*, **281**, 26253–9.
17. Halbedel, S., and Stülke, J. 2007, Tools for the genetic analysis of Mycoplasma. *Int. J. Med. Microbiol.*, **297**, 37–44.
18. Lyon, B. R., May, J. W., and Skurray, R. A. 1984, Tn4001: a gentamicin and kanamycin resistance transposon in *Staphylococcus aureus*. *Mol. Gen. Genet.*, **193**, 554–6.
19. Dybvig, K., French, C. T., and Voelker, L. L. 2000, Construction and use of

derivatives of transposon Tn4001 that function in *Mycoplasma pulmonis* and *Mycoplasma arthritidis*. *J. Bacteriol.*, **182**, 4343–7.

20. Pour-El, I., Adams, C., and Minion, F. C. 2002, Construction of mini-Tn4001tet and its use in *Mycoplasma gallisepticum. Plasmid*, **47**, 129–37.
21. Hahn, T. W., Mothershed, E. A., Waldo, R. H., 3rd, and Krause, D. C. 1999, Construction and analysis of a modified Tn4001 conferring chloramphenicol resistance in *Mycoplasma pneumoniae*. *Plasmid*, **41**, 120–4.
22. Algire, M. A., Lartigue, C., Thomas, D. W., Assad-Garcia, N., Glass, J. I., and Merryman, C. 2009, New selectable marker for manipulating the simple genomes of Mycoplasma species. *Antimicrob. Agents Chemother.*, **53**, 4429–32.
23. Pich, O. Q., Burgos, R., Planell, R., Querol, E., and Piñol, J. 2006, Comparative analysis of antibiotic resistance gene markers in *Mycoplasma genitalium*: application to studies of the minimal gene complement. *Microbiology*, **152**, 519–27.
24. Chopra-Dewasthaly, R., Zimmermann, M., Rosengarten, R., and Citti, C. 2005, First steps towards the genetic manipulation of *Mycoplasma agalactiae* and *Mycoplasma bovis* using the transposon Tn4001mod. *Int. J. Med. Microbiol.*, **294**, 447–53.
25. Beaman, K. D., and Pollack, J. D. 1983, Synthesis of adenylate nucleotides by *Mollicutes* (mycoplasmas). *J. Gen. Microbiol.*, **129**, 3103–10.
26. Jores, J., Fischer, A., Sirand-Pugnet, P., et al. 2013, *Mycoplasma feriruminatoris* sp. nov., a fast growing Mycoplasma species isolated from wild *Caprinae*. *Syst. Appl. Microbiol.*, **36**, 533–8.
27. Gibson, D. G., Young, L., Chuang, R.-Y., Venter, J. C., Hutchison, C. A., 3rd, and Smith, H. O. 2009, Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods*, **6**, 343–5.
28. 1987, The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*
29. Felsenstein, J. 1985, Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, **39**, 783.
30. Tamura, K., Nei, M., and Kumar, S. 2004, Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proc. Natl. Acad. Sci. U. S. A.*, **101**, 11030–5.
31. Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. 2018, MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.*, **35**, 1547–9.
32. Omotajo, D., Tate, T., Cho, H., and Choudhary, M. 2015, Distribution and diversity of ribosome binding sites in prokaryotic genomes. *BMC Genomics*, **16**, 604.
33. Zimmerman, C.-U., -U. Zimmerman, C., and Herrmann, R. 2005, Synthesis of a small, cysteine-rich, 29 amino acids long peptide in *Mycoplasma pneumoniae*. *FEMS Microbiology Letters*, pp. 315–21.
34. Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., et al. 2008, Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–9.
35. Jackman, S. D., Vandervalk, B. P., Mohamadi, H., et al. 2017, ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. *Genome Res.*, **27**, 768–77.

36. Xu, H., Luo, X., Qian, J., et al. 2012, FastUniq: A fast de novo duplicates removal tool for paired short reads. *PLoS One*, **7**, e52249.
37. Bolger, A. M., Lohse, M., and Usadel, B. 2014, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–20.
38. Langmead, B., and Salzberg, S. L. 2012, Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–9.
39. Li, H., Handsaker, B., Wysoker, A., et al. 2009, The sequence alignment/map format and SAM tools. *Bioinformatics*, **25**, 2078–9.
40. Huerta-Cepas, J., Forslund, K., Coelho, L. P., et al. 2017, Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Mol. Biol. Evol.*, **34**, 2115–22.
41. Junier, I., Unal, E. B., Yus, E., Lloréns-Rico, V., and Serrano, L. 2016, Insights into the mechanisms of basal coordination of transcription using a genome-reduced bacterium. *Cell Syst*, **7**, 227–9.
42. Schrader, J. M., and Uhlenbeck, O. C. 2011, Is the sequence-specific binding of aminoacyl-tRNAs by EF-Tu universal among bacteria? *Nucleic Acids Res.*, **39**, 9746–58.
43. Cammarano, P., Tiboni, O., and Sanangelantoni, A. M. 1989, Phylogenetic conservation of antigenic determinants in archaebacterial elongation factors (Tu proteins). *Can. J. Microbiol.*, **35**, 2–10.
44. Yus, E., Yang, J.-S., Sogues, A., and Serrano, L. 2017, A reporter system coupled with high-throughput sequencing unveils key bacterial transcription and translation determinants. *Nat. Commun.*, **8**.
45. Lloréns-Rico, V., Lluch-Senar, M., and Serrano, L. 2015, Distinguishing between productive and abortive promoters using a random forest classifier in *Mycoplasma pneumoniae*. *Nucleic Acids Res.*, **43**, 3442–53.
46. Mariscal, A. M., González-González, L., Querol, E., and Piñol, J. 2016, All-in-one construct for genome engineering using Cre-lox technology. *DNA Res.*, **23**, 263–70.
47. Dybvig, K., Lao, P., Jordan, D. S., and Simmons, W. L. 2010, Fewer essential genes in mycoplasmas than previous studies suggest. *FEMS Microbiol. Lett.*, **311**, 51–5.
48. Baby, V., Lachance, J.-C., Gagnon, J., et al. 2018, Inferring the minimal genome of *Mesoplasma florum* by comparative genomics and transposon mutagenesis. *mSystems*, **3**.
49. Li, H., Xiao, L., Zhang, L., et al. 2018, FSPP: A Tool for genome-wide prediction of smORF-encoded peptides and their functions. *Front. Genet.*, **9**, 96.
50. Miravet-Verde, S., Ferrar, T., Espadas-García, G., et al. 2019, Unraveling the hidden universe of small proteins in bacterial genomes. *Mol. Syst. Biol.*, **15**, e8290.
51. Goryshin, I. Y., and Reznikoff, W. S. 1998, Tn5 in vitro transposition. *J. Biol. Chem.*, **273**, 7367–74.
52. Liu, H., Price, M. N., Waters, R. J., et al. 2018, Magic pools: parallel assessment of transposon delivery vectors in bacteria.

For access all the material available for this article, please visit the following QR.

SCAN ME

3. Chapter 3.

Montero-Blay, A., Piñero-Lambea, C., Miravet-Verde, S., Lluch-Senar, M. & Serrano, L. Inferring Active Metabolic Pathways from Proteomics and Essentiality Data. Cell Rep. 31, 107722 (2020).

Inferring active metabolic pathways from proteomics and essentiality data

Montero-Blay A [1,2], Piñero-Lambea C [1,2], Miravet-Verde S [1,2], Lluch-Senar M [1,2], Serrano L [1,2,3*]

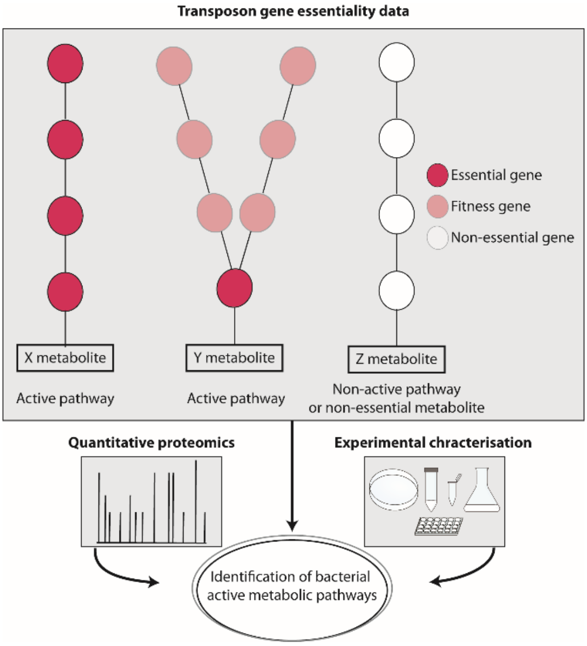
1 EMBL/CRG Systems Biology Research Unit, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003, Spain

2 Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain

3 Institució Catalana de Recerca i Estudis Avançats (ICREA), 08010 Barcelona, Spain

*Corresponding author

Graphical abstract


Transposon gene essentiality data
Essential gene
Fitness gene
Non-essential gene
X metabolite
Y metabolite
Z metabolite
Active pathway
Active pathway
Non-active pathway or non-essential metabolite
Quantitative proteomics
Experimental chracterisation
Identification of bacterial active metabolic pathways

Abstract

During evolution, each bacterial strain shapes its metabolism in order to colonize a diversity of niches. Here we propose an approach to identify active metabolic pathways, by integrating gene essentiality analysis and protein abundance. As an example, we used two bacterial species (*Mycoplasma pneumoniae* and *Mycoplasma agalactia*e) that share a high gene content similarity yet show significant metabolic differences. After integrating all available metabolic knowledge about their enzymes, metabolites and reactions, we built detailed metabolic maps of their carbon metabolism. The most striking difference being the absence of two key enzymes for glucose metabolism in *M. agalactiae*. We determined the carbon sources that allow growth in *M. agalactiae* and we introduced glucose-dependent growth in *M. agalactiae* to show the functionality of its glycolytic enzymes. By analyzing gene essentiality and performing quantitative proteomics, we could predict the active metabolic pathways connected to carbon metabolism and show significant differences in use and direction of key pathways despite sharing the large majority of genes. Comparison between predicted and experimentally determined active pathways shows an excellent agreement. Thus, protein essentiality profiling using transposon sequencing analysis combined with quantitative proteomics and metabolic maps could be used to determine activity and directionality of metabolic pathways.

Introduction

Evolution has shaped bacterial metabolism during millions of years, allowing these microorganisms to successfully colonize an enormous variety of environments, hosts, or tissues within a particular host (Rottem, 2003). Aside from its interest from an evolutionary point of view, understanding bacterial metabolism is key for a wide variety of applications. For instance, bacteria are attractive workhorses for the green production of valuable compounds, such as those related with the denim industry, that have been classically produced using environmentally hazardous processes (Hsu *et al*, 2018). In addition, studying the metabolism of a bacterium is essential to understand its interaction with the bacterium's host. Specifically, it has recently been shown that the microbiome metabolism can affect dramatically drug pharmacokinetics (Zimmermann *et al*, 2019), which therefore should be taken into account for personalized medicine purposes. Lastly, a detailed knowledge of all chemical reactions that take place in a particular strain could help to develop attenuated vaccines or bactericidal therapies based on selectively toxic metabolic intermediates.

Understanding metabolism first requires being able to build an accurate metabolic map with no dead-end reactions or futile loops. This could be hindered by the fact that some metabolic enzymes can use different substrates and produce different products. For instance, fructose bisphosphate aldolase (Fba) is a key enzyme of the glycolytic pathway that catalyzes the conversion of fructose 1,6-bisphosphate (fructose-1,6BP) to dihydroxyacetone phosphate (DHAP) and glyceraldehyde phosphate (GAP). However, the same enzyme can also perform other less well-known reactions, such as conversion of fructose 1-phosphate (fructose-1P) to DHAP and glyceraldehyde (GA), or conversion of sedoheptulose 1,7-bisphosphate (sedoheptulose-1,7BP) to DHAP and erythrose 4-phosphate (erythrose-4P), among others. This catalytic versatility of metabolic enzymes could explain why in the experimentally validated metabolic reconstruction performed in *M. pneumoniae* (Yus *et al*, 2009), some of the predicted reactions were not supported by any known enzymatic activity encoded by *M. pneumoniae* genome.

Here, we addressed whether it is possible to discern active and important metabolic pathways using available omics data, such as proteomics and transposon-determined gene essentiality. The basis for this hypothesis is that, in principle, pathways that produce essential metabolites for the cell must be composed of enzymes that are either essential (e.g. cannot be deleted or disrupted) or necessary for fitness (e.g. disruption triggers a growth impairment). In other words, if the pathway leading to an essential metabolite is unique, all enzymes involved should be essential; in contrast, if there is an alternative pathway with less capacity or whose use could detract from making another important metabolite, the enzymes involved should be for fitness. Enzymes involved in pathways that are not actively used under certain conditions or that generate non-essential metabolites will be non-essential, and their deletion should not compromise bacterial growth. Furthermore, we could imagine that, in general proteins in a pathway with no branches should exhibit less variation in terms of protein abundance, while proteins upstream or downstream of a branching point will have different concentration than those in the linear pathway and therefore more variability in terms of protein abundance. Depending on the activity of the different branches, protein concentrations will be different for each branch. By combining both concepts, we potentially would be able to predict the activity and importance of metabolic active pathways. This information could be used in metabolic flux analysis or bacterial engineering. In fact the use of transposon insertional profile analysis have been used in other bacteria for identifying specific metabolic genes in a determined condition (Ochsner *et al,* 2017) or as a tool to validate existing genomic-based metabolic reconstruction and flux balance analysis (Yang *et al*, 2014).

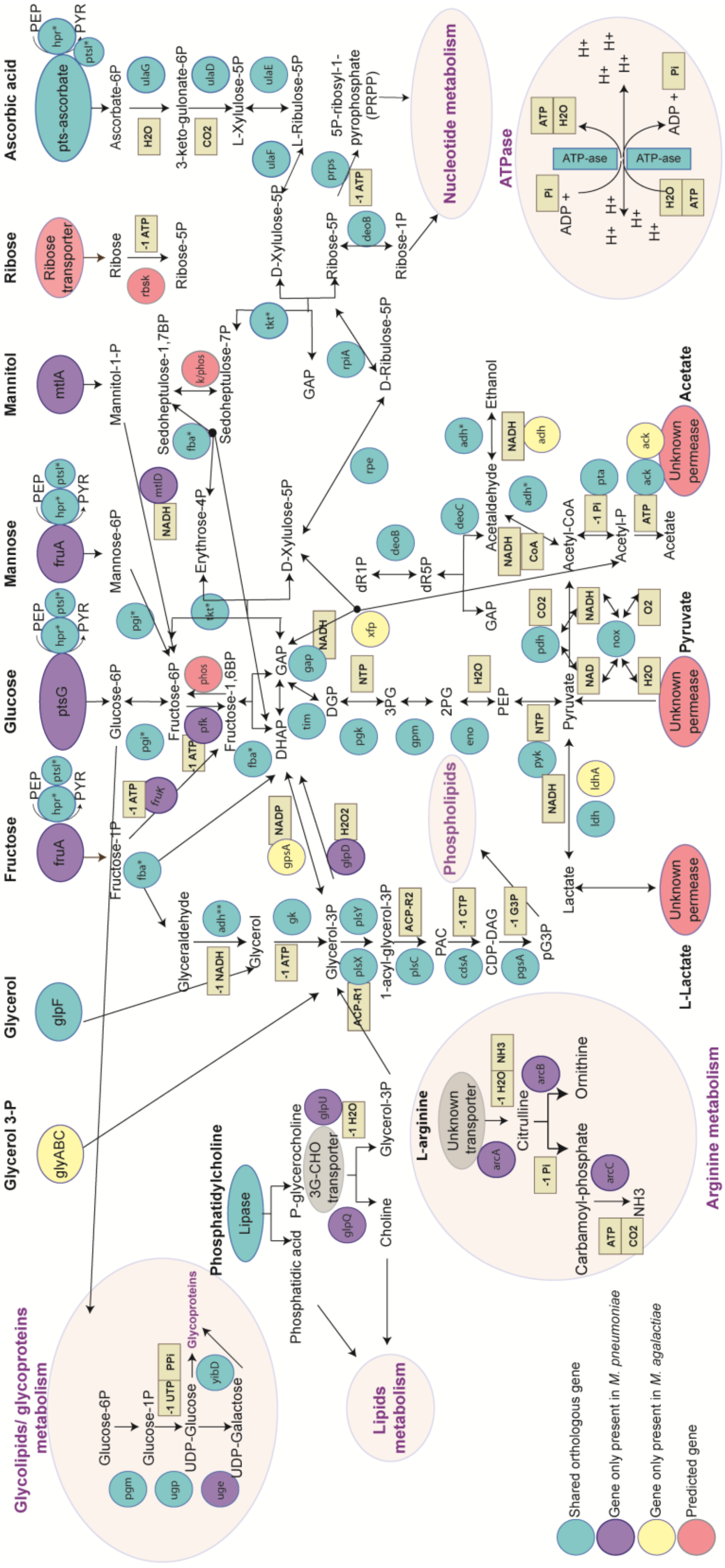

Ascorbic acid
pts-ascorbate
hpr*
ptsI*
PEP
PYR
Ascorbate-6P
ulaG
H2O
3-keto-gulonate-6P
ulaD
CO2
L-Xylulose-5P
ulaE
L-Ribulose-5P
ulaF
D-Xylulose-5P
prps
-1 ATP
5P-ribosyl-1-pyrophosphate (PRPP)
Nucleotide metabolism
ATPase
ATP
H2O
ATP-ase
ATP-ase
ADP +
Pi
H+
Ribose
Ribose transporter
Ribose
rbsk
-1 ATP
Ribose-5P
deoB
Ribose-1P
tkt*
GAP
rpiA
D-Ribulose-5P
Mannitol
mtlA
Mannitol-1-P
k/phos
Sedoheptulose-1,7BP
Sedoheptulose-7P
fba*
mtlD
NADH
rpe
Mannose
fruA
hpr*
ptsI*
Mannose-6P
Erythrose-4P
D-Xylulose-5P
adh*
NADH
adh
Ethanol
Acetaldehyde
adh*
deoC
NADH
CoA
Acetyl-CoA
pta
-1 Pi
Acetyl-P
ack
ack
ATP
Acetate
Unknown permease
Acetate
deoB
dR1P
dR5P
GAP
Glucose
ptsG
Glucose-6P
pgi*
pgi*
tkt*
Fructose-6P
phos
pfk
-1 ATP
Fructose-1,6BP
fba*
DHAP
tim
GAP
gap
NADH
xfp
DGP
NTP
pgk
3PG
gpm
2PG
H2O
eno
PEP
NTP
pyk
Pyruvate
pdh
CO2
NADH
NAD
nox
O2
H2O
Unknown permease
Pyruvate
Fructose
fruA
fruK
-1 ATP
Fructose-1P
fba*
NADH
ldh
ldhA
Lactate
Unknown permease
L-Lactate
NADP
gpsA
H2O2
glpD
Phospholipids
Glycerol
glpF
Glyceraldehyde
adh**
-1 NADH
Glycerol
gk
-1 ATP
Glycerol-3P
plsY
plsX
ACP-R1
1-acyl-glycerol-3P
ACP-R2
plsC
PAC
-1 CTP
cdsA
CDP-DAG
-1 G3P
pgsA
pG3P
Glycerol 3-P
glyABC
Phosphatidylcholine
Lipase
Phosphatidic acid
P-glycerocholine
glpU
3G-CHO transporter
-1 H2O
glpQ
Glycerol-3P
Choline
L-arginine
Unknown transporter
arcA
Citrulline
-1 H2O
NH3
arcB
Ornithine
-1 Pi
Carbamoyl-phosphate
arcC
ATP
CO2
NH3
Arginine metabolism
Lipids metabolism
Glycolipids/ glycoproteins metabolism
pgm
Glucose-6P
Glucose-1P
ugp
-1 UTP
PPi
UDP-Glucose
Glycoproteins
yibD
uge
UDP-Galactose
Shared orthologous gene
Gene only present in M. pneumoniae
Gene only present in M. agalactiae
Predicted gene

Figure 1. Genetic context for selected metabolic genes in M. agalactiae and M. pneumoniae. *Panel showing genes involved in glycolipids, glycoproteins, arginine, and sugar metabolism, and their link to generate DNA or RNA precursors. Blue circles indicate orthologous genes present in both species, purple circles, genes exclusively present in M. pneumoniae, and yellow circles, genes present only in M. agalactiae. Pink indicates assigned genes for which the gene assignation is not known. Single asterisks (*) indicate protein-coding genes participating in multiple reactions in the cell, and double asterisks (**), protein-coding genes described to have multiple functionalities but for which the logics of the pathway architecture (i.e. the absence of the antecedent metabolite) hampers the existence of this specific function. Arrows connecting different metabolites indicate directionality of chemical reactions; molecules or cofactors involved in the different reactions are highlighted in squared boxes.*

To see whether active metabolic pathways could be inferred from proteomics and essentiality data, we selected two bacterial species in the Mycoplasma genus: *Mycoplasma pneumoniae* and *Mycoplasma agalactiae*, which comprises species with streamlined genomes and simplified metabolic complexity. For *M. pneumoniae*, a species that infects the human lung and causes atypical pneumonia (Waites & Talkington, 2004), there is a high-resolution gene essentiality study (Lluch-Senar *et al*, 2015), metabolic reconstruction (Yus *et al*, 2009), measurement of metabolites and fluxes (Maier *et al*, 2013), metabolic modeling and flux balance analysis (FBA) (Wodke *et al*, 2013). In the case of *M. agalactiae,* a metabolic analysis with isotope tracking using different carbon sources is available for the closely related species *Mycoplasma bovis (*Masukagami *et al*, 2017). *M. bovis* and *M. agalactiae* share 89.6% of their genes (Qi *et al*, 2012) and have a 99% sequence similarity in their 16S rRNA (Pettersson *et al*, 1996). Furthermore, for each of the genes involved in carbohydrates, lipids, amino acids, nucleotides and vitamins metabolism in *M. agalactiae,* it has been found an orthologous gene in the *M. bovis* genome evidencing their metabolic analogy (Supplementary Table 1). For both bacteria there are high-resolution essentiality studies available (Lluch-Senar *et al*, 2015 and Montero-Blay *et al*, 2019).

Here, we first reconstructed the metabolic pathways using genome and metabolite information available. Then, we experimentally determined the carbon sources that allowed robust growth and determined protein relative concentration by mass spectroscopy. Finally we mapped gene essentiality on the metabolic maps. Comparisons of the metabolic active pathways predicted by our omics analysis with those found in the literature for both bacterial species showed an excellent agreement.

Results

Generation of the metabolic maps

We first constructed metabolic maps with all possible enzymes involved in carbon metabolism that are capable of producing ATP, as well as with the direct branches coming out of them, into nucleotides, lipids, glycolipids, and glycoproteins for *M. pneumoniae* and *M. agalactiae* (Figure 1).
For the generation of the metabolic map we used genomics data from *M. pneumoniae* (Himmelreich *et al,* 1996) and *M. agalactiae* (Montero-*Blay et al,* 2019). This information was consistent with KEGG database (Kanehisa and Goto 2000). Though there are computational methods to fulfil missing reactions or existing metabolic gaps (Karp *et al*, 2018 and Ponce-de-Leon *et al*, 2019), the simplicity of Mycoplasmas with around 700 genes allowed us to manually curate the metabolic pathways using the maps for *M. pneumoniae* of Yus *et al*, 2009 and Wodke *et al*, 2013 as a reference. A table for orthologous genes from *M. pneumoniae, M. agalactiae* and *M. bovis* has been generated in this work (Supplementary Table 1).

Comparative genomics reveals fundamental differences in *M. agalactiae* and *M. pneumoniae* carbon metabolism

At first glance, *M. agalactiae* and *M. pneumoniae* appear to share a common core of metabolic reactions (Figure 1, blue circles). However, a more detailed look reveals that *M. agalactiae* has lost some of the pathways involved in carbon metabolism during evolution that are present in *M. pneumoniae* (Figure 1, purple circles). Likewise, there are certain enzymatic activities missing in *M. pneumoniae*, which are only encoded by *M. agalactiae* genome (Figure 1, yellow circles).

M. agalactiae lacks genes encoding transporters involved in the metabolism of glucose (*ptsG*), fructose and mannose (*fruA*), mannitol (*mtlA*), and glycerophosphocholine (*glpU*). Accordingly with this lack of transporters, all the cytoplasmic processing enzymes whose activity is exclusively involved in the metabolism of these carbon sources are also absent in the genome of *M. agalactiae* (i.e. *fruK* for fructose, *mtlD* for mannitol and *glpQ* for glycerophosphocholine). The only exception to this is the maintenance of cytoplasmic enzymes involved in glucose and mannose metabolism that is discussed below.

Regarding carbon metabolism, *M. agalactiae* has some unique features. First, as indicated above, it lacks a PTS-glucose transporter as well as phosphofructokinase (Pfk), and can therefore only enter into glycolysis from the phosphate pentose pathway (PPP) or using glycerol-3P. Second, whereas the genes encoding lactate dehydrogenase, alcohol dehydrogenase, and acetate kinase are present in a single copy in *M. pneumoniae* genome (i.e. *MPN674*, *MPN564,* and *MPN533*, respectively), these same genes are duplicated in the *M. agalactiae* genome (*MAGA7784*_RS00770 / *MAGA7784_RS02675, MAGA7784_RS02235*A / *MAGA7784_RS02235*B, and

MAGA7784_RS00725 / *MAGA7784_RS02690* respectively). Third, *M. agalactiae* contains unique genes not present in *M. pneumoniae* (*xfp* and *gpsA*). The enzyme D-xylulose 5-phosphate/ D-fructose 6-phosphate phosphoketolase (Xfp) reversibly catalyzes the conversion of D-xylulose to GAP and acetyl-P. This reaction, which is not present in *M. pneumoniae*, connects the PPP to glycolysis and has been detected in a vast range of organisms, from bacteria to yeast (BRENDA: EC4.1.2.9). The GpsA enzyme converts reversibly glycerol-3P into DHAP, using NADPH as a reducing cofactor. In *M. pneumoniae*, the reaction connecting glycerol metabolism and glycolysis is irreversible and is catalyzed by GlpD, producing H_2O_2.

For arginine metabolism, there are genomic but no functional differences between the species: *M. agalactiae* does not carry any gene involved in arginine metabolism, while *M. pneumoniae* contains three genes (i.e. *arcA*, *arcB*, and *arcC*) whose protein products could metabolize arginine (to obtain one molecule of ATP). However, whereas this pathway is fully active in other *Mycoplasma* species (such as *Mycoplasma fermentans*), *M. pneumoniae* has lost its ability to metabolize arginine, as indicated by the presence of severe truncations in the genes encoding enzymes of this pathway (Rechnitzer *et al*, 2013).

For glycolipid metabolism, both species contain a functional pathway to produce UDP-glucose from glucose-6P. However, the conversion of UDP-glucose to UDP-galactose is only enzymatically supported in *M. pneumoniae*, but not in *M. agalactiae*. Both metabolites can be used to generate glycoproteins as well. In line with this, it has been experimentally demonstrated that the levels of UDP-glucose are much higher in the *M. agalactiae* counterpart, *M. bovis*, than in the *M. pneumoniae*–related species *M. gallisepticum* (Masukagami *et al*, 2017).

Regarding phospholipid production, both organisms share the same set of genes. However, *M. pneumoniae* can only import glycerol through a permease (i.e. GlpF), as the predicted ABC transporter for glycerol 3-P encoded by its genome has been shown to carry out other functions Großhennig *et al*, 2013). In contrast, aside from the GlpF permease, *M. agalactiae* genome also encodes a dedicated ABC transporter for glycerol 3-P, that in contrast with the one predicted for *M. pneumoniae*, shares a high homology with the validated glycerol 3-P uptake system form *Escherichia coli*.

Filling gaps in the metabolic map

There are genes present in both species (*fba* and *pgi*) whose maintenance in *M. agalactiae* is difficult to explain, given the absence of key enzymes in the import and metabolism of glucose. Fba catalyzes the reversible split of fructose-1,6BP into DHAP and GAP as part of the glycolytic pathway present in *M. pneumoniae*, and its substrate (i.e. fructose-1,6BP) is produced by Pfk, which is absent in *M. agalactiae*. However, as mentioned in the introduction, Fba can also reversibly catalyze the synthesis of sedoheptulose-1,7BP from DHAP and erythrose-4P (Vanyushkina *et al*, 2014) and therefore participate in the PPP. This less known enzymatic activity of Fba would explain the absence of transaldolase coding gene in both species as well as in other Mycoplasma species (Kamminga *et al,* 2017, Vanyushkina *et al*, 2014). Transaldolase is the enzymatic activity that usually connects PPP and glycolysis in most living forms. Of note, the Fba-mediated connection of PPP and glycolysis requires the presence of a sedoheptulose-7P kinase to convert this metabolite into sedoheptulose-1,7BP, and/or a phosphatase that performs the opposite reaction (see below). These two metabolites have been identified in *M. pneumoniae* (sedoheptulose-7P (Maier *et al*, 2013)) and *M. bovis* (sedoheptulose-7P and sedoheptulose-1,7BP, (Masukagami *et al*, 2017)

(Supplementary Table 2). The existence of both metabolites supports the presence of these two activities.

The reversible enzyme Pgi converts glucose-6P into fructose-6P. This step is isolated from the rest of glycolysis in *M. agalactiae*, as it does not have the Pfk or Pts proteins (Figure 1). On the other hand, there is experimental evidence in *M. bovis* that ^{13}C-glycerol can be converted into fructose-6P (Masukagami *et al*, 2017). This suggests the existence of a phosphatase that converts fructose-1,6BP into fructose-6P. The metabolic product of this phosphatase activity would then be converted into glucose-6-P, which enters into the glycolipids and glycoproteins pathway that leads to essential cell metabolites. In *E. coli*, different phosphatases (i.e. YieH, YbiV, YidA and YaeD) could dephosphorylate fructose-1,6BP, into fructose-6P. Of these, YidA is the enzyme with lowest K_m (Kuznetsova *et al*, 2006). *M. agalactiae* has different sugar phosphatases (e.g. *MAGA7784_RS01220, MAGA7784_RS00145; MAGA7784_RS03930*), and one of these is related to YidA in *E.coli* (*MAGA7784_RS03930*) (Supplementary Figure 1). Thus, we propose that *MAGA7784_RS03930* could catalyze the conversion of fructose-1,6BP into fructose-6P.

Experimental validation of carbon sources capable of supporting growth in *M. agalactiae*

For *M. pneumoniae*, the carbon sources capable of promoting growth have been previously analyzed (Yus *et al*, 2009). This bacterium grows best when using glucose as carbon source, but is also able to grow with mannose, fructose, ribose, glycerol, ascorbate (Yus *et al*, 2009), and phosphoglycerol choline (Großhennig *et al*, 2013), although to a much lesser extent. However, it cannot use pyruvate, lactate, mannitol, or arginine.

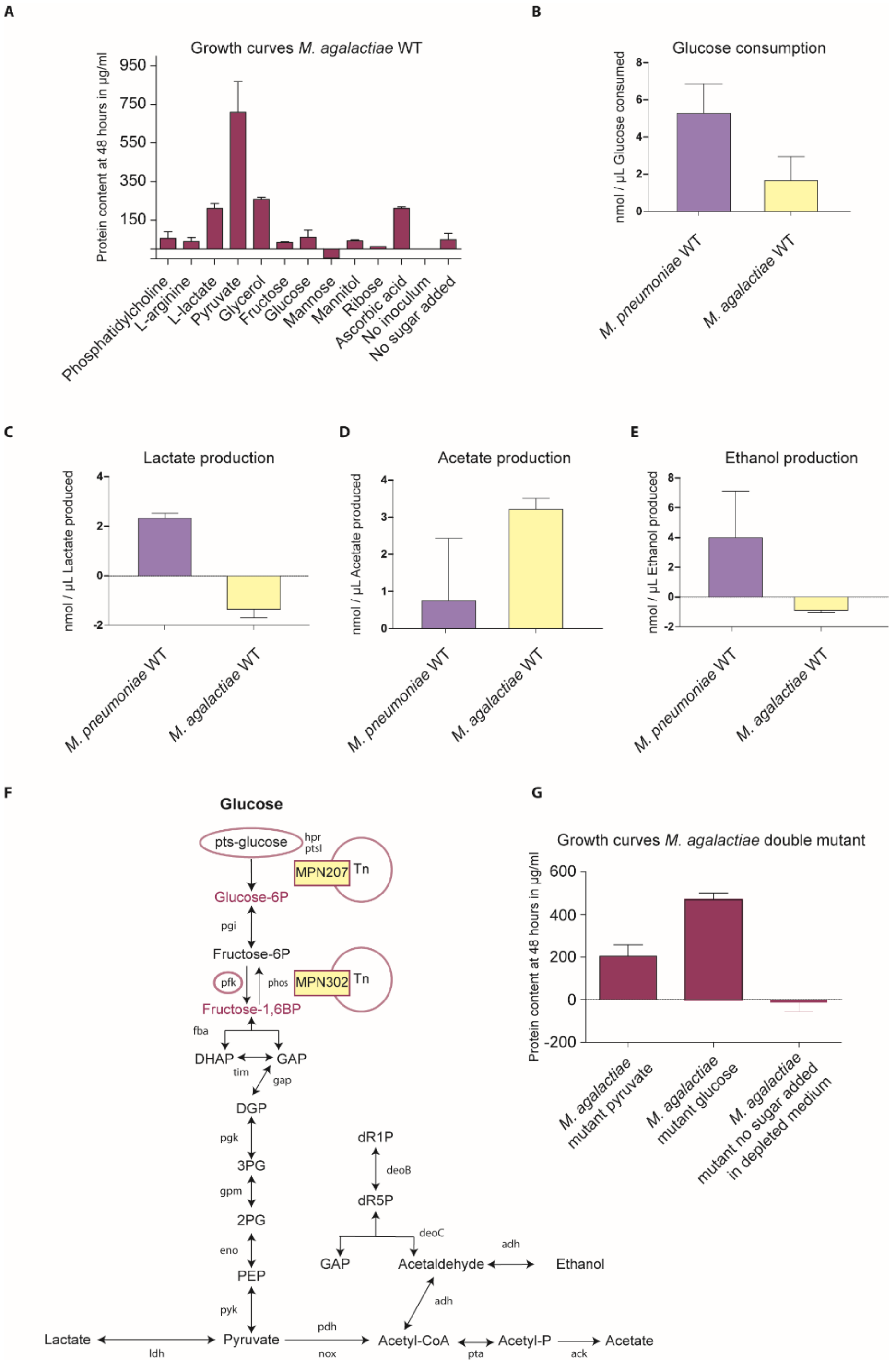


***Figure 2. Growth measurement of M. agalactiae under different culture conditions, and secondary metabolites measurement for M. agalactiae and M. pneumoniae**. A Measurement of protein content after 48 h of growth for M. agalactiae cultures, under various culture conditions using different carbon sources in depleted medium. B, C, D, E Measurement of glucose consumption or lactate, acetate, or ethanol production in M. pneumoniae (left) or M. agalactiae (right). F Reconstruction of the glycolysis pathway and pyruvate metabolism for M. agalactiae is indicated by black letters. Genes introduced via transposon from M. pneumoniae, to restore complete glycolytic functionality in M. agalactiae, are indicated by pink letters. G Measurement of protein content after 48 h of growth for the synthetic strain of M. agalactiae that has restored glycolysis when using pyruvate or glucose as a carbon source.*

One initial obstacle for determining the carbon sources required for *M. agalactiae* was our lack of a chemically-defined medium in which the compounds present could be controlled. Notably, Mycoplasmas are characterized by their reduced metabolic and biosynthetic capacities. For this reason, the broth used for their growth, termed Hayflick medium, contains animal-derived serum with unspecified composition, complemented by the addition of the preferred carbon source for each *Mycoplasma* species. The presence of serum in the broth likely explains the modest growth observed for *M. agalactiae* in plain Hayflick medium without any added carbon source (Supplementary Table 3). This limited growth could mask the effect of exogenous added carbon sources. To prevent this, we first generated a “depleted” version of the rich medium in which all carbohydrate likely provided by the serum have been exhausted (see Materials and Methods section, Supplementary Table 3). We then supplemented this depleted medium with 11 different carbon sources, including phosphatidylcholine, L-arginine, L-lactate, pyruvate, glycerol, fructose, glucose, mannose, mannitol, ribose and ascorbic acid, and then assessed the growth of *M. agalactiae*.

No significant growth of *M. agalactiae* was observed in the depleted medium (Figure 2A). As expected based on the metabolic map (Figure 1), we observed significant growth when the depleted medium was supplemented with ascorbic acid or glycerol (Figure 2A). Supplementing with those sugars requiring metabolic enzymes or transporters absent from the *M. agalactiae* genome, such as phosphatidylcholine arginine, ribose, glucose, fructose, or mannose, did not lead to significant growth (Figure 2A). We found that lactate and especially pyruvate supported robust growth of the bacterium (Figure 2A, Supplementary Table 3). This is surprising since these metabolites do not support growth in *M. pneumoniae* and suggests the presence of specific transporters for these metabolites in *M. agalactiae* and the existence of a reverse gluconeogenic flow. Only a gluconeogenic flow starting from pyruvate could provide the required precursors for phospholipids and nucleotide metabolism in a depleted medium supplemented with pyruvate.

In agreement with the metabolic maps, glucose is consumed by *M. pneumoniae* but not by *M. agalactiae* (Figure 2B). This is in line with the observation that *M. bovis*, a species closely related to *M. agalactiae*, cannot metabolize glucose (Masukagami *et al*, 2017). *M. pneumoniae* produces lactate as sub-product of its metabolism (Yus *et al*, 2009), while *M. agalactiae* grown in medium with pyruvate seems to consume the traces of lactate present in the non-depleted version of the medium (Figure 2C). Notably, while both species produced acetate, the amount of acetate produced by *M. agalactiae* was four times that of the concentration measured for *M. pneumoniae* (3.2 nmol/µl for *M. agalactiae* vs 0.8 nmol/µl for *M. pneumoniae*, Figure 2D. Finally, it seems that some ethanol could be produced in *M. pneumoniae* (Figure 2E).

Overall, when metabolizing glucose, *M. pneumoniae* generated the metabolic sub-products of lactate, acetate, and possibly, some ethanol (Supplementary table 4). In contrast, *M. agalactiae* grown in presence of pyruvate mainly generated acetate as sub-product, and lactate was consumed from the medium. These results indicated that while *M. pneumoniae* used glycolysis for ATP production (2 ATP per molecule or 4 ATP per molecule, depending of type of sub-product*;* lactate or acetate, respectively), *M. agalactiae* converted lactate, pyruvate and/or glycerol into acetate to generate only one ATP per molecule. It would also produce two ATPs per molecule when using ascorbate, to produce acetate.

Generation of a synthetic *M. agalactiae* strain capable of metabolizing glucose

Glycolysis is an energy conversion pathway used for many organisms to convert glucose to pyruvate. Glucose is the main carbon source used for *Mycoplasmas*, but the *M. agalactiae* and *M. bovis* species are unable to use glucose for ATP production. Strikingly, despite being unable to metabolize glucose, *M. agalactiae* and *M. bovis* still conserve eight out of the ten genes involved in the glycolytic pathway. We wondered if these remaining genes were still functional in glycolysis or could be involved in other functionalities. To address this point, we decided to complement the *M. agalactiae* genome with the two genes involved in the glycolytic pathway missing in these species (i.e. PTS-glucose transporter and Pfk proteins). These two proteins are encoded by the genes *MPN207* and *MPN302*, respectively, in *M. pneumoniae* (Figure 2F). We cloned the genes either together or individually in a transposon under the control of a strong native regulatory region of *M. agalactiae* (see construct information in Supplementary Table 5) and transformed these constructs into *M. agalactiae.* The expression of both heterologous proteins was verified by mass spectrometry (Supplementary Table 6).

We assessed the ability of the engineered *M. agalactiae* strains to grow in depleted medium supplemented with glucose as the carbon source. Whereas the single mutants were unable to grow, as inferred from their inability to increase the culture biomass (Supplementary Table 3), the synthetic strain carrying copies of both *MPN207* and *MPN302* grew even better than in depleted medium supplemented with pyruvate (Figure 2G). Moreover, this synthetic strain consumed glucose and produced acetate/lactate (Supplementary Table 4, glucose consumption). Altogether, these results demonstrated that the glycolytic enzymes present in *M. agalactiae* have retained their functionality and could work in both glycolysis and gluconeogenesis senses.

Proteomics abundance quantification correlates with metabolic pathway activity

As stated in the introduction we wanted to find out if integration of protein copy numbers into a metabolic map could give clues about the magnitude of metabolic fluxes. To this end, we performed free-label mass spectrometry analyses of *M. agalactiae* and *M. pneumoniae* to obtain a quantitative proteome overview of both species (Supplementary Table 1 and 7).

For visualization, we show the relative protein abundance following a color code based on the log2 normalized average area of the three top peptides of each metabolic enzyme for both species (see Methods section, Figure 3 and Supplementary Table 1 for orthologous gene protein comparison). Alcohol dehydrogenase (Adh) and lactate dehydrogenase (Ldh) were among the proteins whose differences in protein abundance were higher in *M. agalactiae* as compared to their orthologous counterparts in *M. pneumoniae*. *M. agalactiae* carries a gene duplication for Adh, and one of the isoforms is more highly expressed in this bacterium than in *M. pneumoniae* (i.e. 29.0 log2 vs 22.3 log2, respectively). Ldh also has two gene copies in *M. agalactiae* coding for two different isoforms, named Ldh and LdhA. LdhA showed a moderate expression level (i.e. 28.8 log2) and has limited amino acid identity with Ldh from *M. pneumoniae*. The Ldh protein from both species share a high amino acid identity and therefore are likely performing the same role, but their expression levels are quite different (i.e. hereinafter referred to as those protein levels whose difference is > 1 log2) in the two species (i.e. 32.6 log2 for *M. agalactiae* vs. 28.7 log2 for *M. pneumoniae*, 3.89 log2 fold change with p-value=0.01).

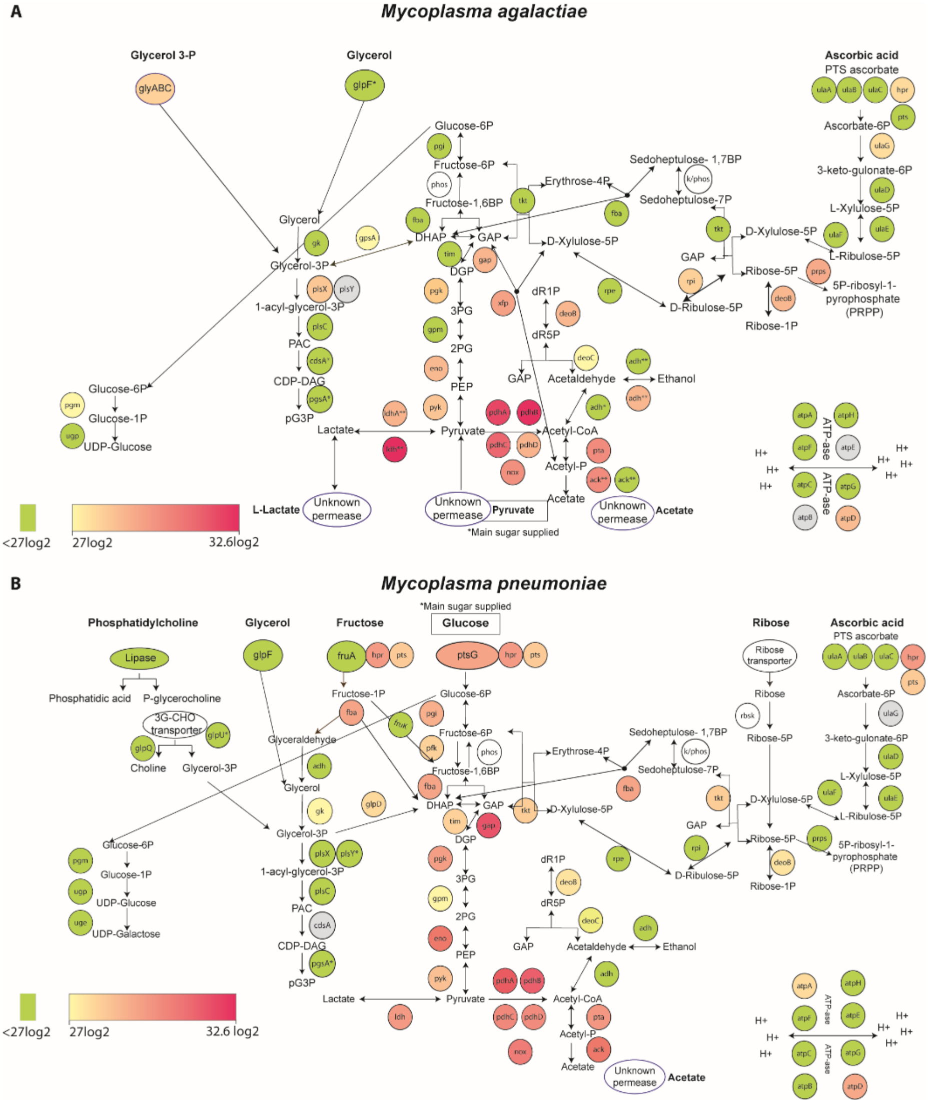


Figure 3. Proteomics abundance for selected metabolic pathways in M. agalactiae and M. pneumoniae. A, B *Normalized protein abundance representation in urea samples for M. agalactiae growing in pyruvate (**A**) or M. pneumoniae growing in*

*glucose (**B**). The color scale represents a gradient according to protein abundance in the set of metabolic enzymes selected, showing a range of minimum (yellow) to maximum (magenta). Proteins with expression levels lower than 27 log2 are shown as min and represented in green. Double asterisks (**) indicates duplicated genes in M. agalactiae. White circles indicate protein-coding genes for proteins with unassigned genes, and grey circles, undetected proteins by mass spectrometry. A single asterisk (*) indicates proteins not detected in proteomics samples using urea as lysis buffer but rescued from SDS samples.*

In addition, there are certain proteins whose relative expression levels are higher in *M. pneumoniae* than in *M. agalactiae.* For instance (and as expected), proteins involved in the glycolytic pathway present in both species (i.e. Pgi, Fba, Tim, Gap, Pgk, Gpm, Eno, and Pyk) are predominantly expressed at higher levels in *M. pneumoniae* than in *M. agalactiae*. This is especially evident for, those enzymes placed in the upper part of the glycolytic pathway (i.e. Pgi, Fba, Tim, and Gap) in which the differences in their relative abundances between the two species are remarkable (i.e. 3.2 log2 fold change p-value=0.04, 3.7 log2 fold change p-value=0.005, 1.9 log2 fold change p-value=0.1, and 2.1 log2 fold change p-value=0.6, respectively). This indicates that, for *M. pneumoniae*, there is a continuous flow from glucose to pyruvate, but that for *M. agalactiae*, there is a break at the level of Tim, in agreement with its incapability to use glucose (see discussion).

There are two other proteins related with glycolysis whose relative expression levels are higher in *M. pneumoniae* than in *M. agalactiae*: PtsI and Hpr. These proteins are involved in the internalization and subsequent phosphorylation of different carbon sources, such as ascorbate, glucose, mannose, and fructose. Whereas both Hpr and PtsI were highly expressed in *M. pneumoniae* (29.7 log2 and 28.2 log2, respectively), they were expressed at lower levels in *M. agalactiae* (27.7 log2 and 24.1 log2, respectively and 2.0 log2 fold change p-value= 0.001; 4.1 log2 fold change p-value=0.09 for Hpr and PtsI). This observation can be explained by the fact that *M. pneumoniae* has more carbon-source transporters that require the use of Hpr and PtsI than *M. agalactiae*. For the PPP, the expression of the transketolase (Tkt) enzyme was also higher in *M. pneumoniae* than in *M. agalactiae* (28.2 log2 vs. 24.8 log2, respectively). This enzyme not only catalyzes an intermediate reaction in the PPP but is also the unique entry point to the PPP from glycolysis in *M. pneumoniae*; in contrast, *M. agalactiae* can also connect to the PPP through the more abundant Xfp enzyme. For ascorbate, both species showed lower but close to similar (hereinafter referred to as those differences ≤ 1 log2) abundances (27.8 log2 for *M. pneumoniae* and 26.7 log2 for *M. agalactiae*), suggesting that this metabolite has a subsidiary role as a carbon source.

Finally, for the metabolism of glycerol/phospholipids as well as glycolipids/glycoproteins, we see similar low levels (e.g. for GlpK enzyme 27.3 log2 and 26.5 log2 for *M. pneumoniae* or *M. agalactiae* respectively) for the enzymes involved in both species. In summary, for *M. agalactiae*, higher expression levels are found in those enzymes involved in pyruvate and lactate metabolism, while for *M. pneumoniae*, higher expression levels are found in general for all glycolytic enzymes and proteins involved in carbohydrate import.

As a general observation derived from the proteomics analysis, proteins belonging to a section of metabolic pathways without bifurcation points (hereinafter referred to as sub-pathways) tend to have similar expression levels. To further demonstrate this, we defined a set of linear sub-pathways for both *M. pneumoniae* and *M. agalactiae* (see methods). Subsequently, the coefficients of variation (CV) for protein abundance were calculated for each individual sub-pathway (observed) as well as for artificial sub-pathways whose members were randomly sampled from the set of metabolic genes of each strain (expected) (Supplementary Table 8 and Supplementary Figure 2). In general, most of the observed sub-pathways showed less CV than expected. However, some exceptions to this trend are observed. For instance, sub-pathway two in *M. pneumoniae* showed a CV similar to the expected if genes were randomly sampled. This might be related with the fact that genes belonging to this sub-pathway are expressed at low levels (i.e. close to the detection limit), which could impair accurate protein abundance determination. Of note, despite these exceptions, when the CV of all the sub-pathways analyzed for *M. pneumoniae* were pooled together, the mean of them was significantly lower than the one observed for artificial sub-pathways for natural quantification or $\log_2$ transformed values respectively (Figure 4A and 4B). Moreover, a similar result was obtained in the *M. agalactiae* analysis (Figure 4C and 4D). These results support the idea that linear sub-pathways have less variation in protein abundance than expected by chance. Furthermore, this correlation in protein abundance cannot be explained by operon composition, as numerous members of the analyzed sub-pathways belong to different transcriptional units (Supplementary Table 9). Therefore, similar protein expression levels in consecutive proteins might be taken as an indication of a functional pathway lacking bifurcation points. Strikingly, this correlation in protein abundance for enzymes belonging to the same sub-pathway seems to be a dynamic phenomenon. In fact, *M. pneumoniae* naturally expresses low levels of enzymes related with fructose metabolism and consequently grows poorly when this sugar is added as major carbon source. However, adaptation to this sugar (i.e. several passages with fructose as main carbon source present in the medium) results in a significant increase of the expression levels of those genes involved in fructose metabolism (Yus *et al*, 2009, Supplementary Figure 3).

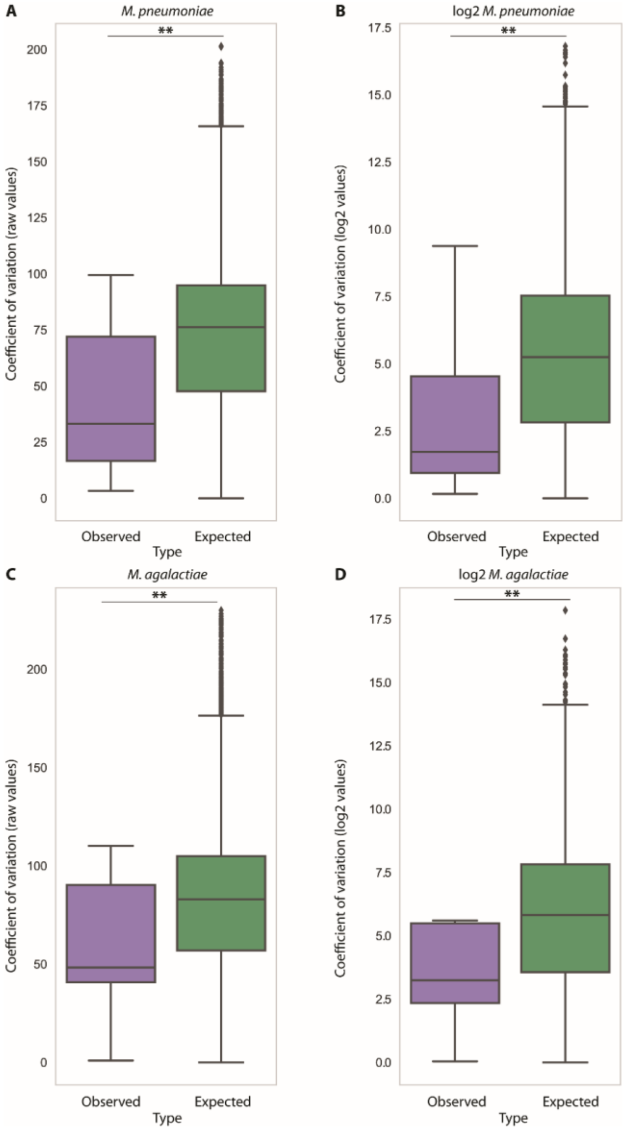


Figure 4. Study of the protein variability in linear pathways. *For each of the strains the coefficient of variation (CV, standard deviation corrected by the mean) for linear sub-pathways was calculated and pooled together. In box-plot* ***A*** *and* ***B*** *for M. pneumoniae and* ***C*** *and* ***D*** *for M. agalactiae using raw quantification values and a log$_2$ transformation respectively. In purple is represented the pooled CV from the set of sub-pathways and in green, the pooled expected CV calculated based on artificial sub-pathways whose members were randomly sampled from the set of metabolic genes of each strain. The P-values obtained for M. pneumoniae are p-value=0.005 and p-value=0.012 for raw and log2 values, respectively. For M. agalactiae, the p-*

values obtained are for raw and log2 values respectively, p-value=0.045 and p-value=0.008

High-resolution transposon density and its application for flux metabolic directionality ascertainment

We propose that gene essentiality could provide insights about the activity of different metabolic pathways of a cell, which together with information about protein abundance would complement the information provided by metabolomics studies. The assumption behind this is that the level of essentiality of a gene will be related to the importance of the pathway and the flux that goes through it. However, this exclusively applies for essential and fitness genes whilst for non-essential genes the flux going through it does not necessarily have to be zero.

To address this, we plotted the metabolic maps of *M. agalactiae* and *M. pneumoniae* together with a color-coded scale of transposon insertion density (see Methods, Figure 5A and Figure 5B, and Supplementary Table 10). Insertion density is related to the impact that disruption of the gene has on the fitness of the bacterium (Lluch-Senar *et al*, 2015; Montero-Blay *et al*, 2019). Our essentiality/metabolic map in *M. agalactiae* reflected an essential footprint along all the genes involved in converting pyruvate into acetate (i.e. the four subunits of the pyruvate dehydrogenase complex, NADH oxidase, phosphate acetyltransferase, and acetate kinase). Out of the two copies, the more abundantly expressed *ack* gene showed an essential profile (shown in magenta), whereas the other less-abundant isoform was clearly non-essential (shown in white). This divergence suggests that these two isoforms have different functionalities, as if they were performing the same role it would impossible to find an essential character for any of the isoforms. Conversely, the lactate related genes *ldh* and *ldhA* were found as fitness under these conditions. This falls in line with the fact that *M. agalactiae* in order to generate ATP produces acetate rather than lactate as main sub-product of pyruvate metabolism. Unfortunately, the dispensability of the lactate pathway in *M. agalactiae* prevents to ascertain whether the two isoforms Ldh and LdhA have the same function. In contrast, in *M. pneumoniae*, the entire pathway from pyruvate to acetate showed a fitness footprint, as did the Ldh enzyme. This is expected, as *M. pneumoniae* glycolysis produces two ATPs before bifurcation into lactate or acetate and therefore does not need the ATP produced when going from pyruvate to acetate.

Even though *M. agalactiae* is unable to use glucose as a carbon source, the genes *pyk*, *eno*, *gpm*, *pgk*, and *gap* involved in the lower part of the glycolytic pathway showed an essential profile. This was not the case for the Fba and Pgi enzymes (discussed above). As the essentiality map was performed with pyruvate as the main carbon source, this suggests the existence of a gluconeogenic flux (i.e. reverse glycolysis) from pyruvate to GAP in this bacterium. The essential footprint extended across most of the genes (i.e. *xfp*, *rpi*, *rpe*, *deoB*, and *prps*) involved in the conversion of GAP to 5-phospho-D-ribosyl α-1-pyrophosphate (PRPP), ribose-5P, and ribose-1P, all of which are critical precursors for DNA and RNA metabolism. The *xfp* gene shows a fitness profile that can be explained, as the role of Xfp converting GAP into xylulose 5-P can also be mediated by Tkt, and further agrees with the fact that we did not observe growth of *M. agalactiae* with ribose. In *M. pneumoniae*, the entry into ribose-5P and ribose-1P could be direct, as it can grow on ribose (as shown its presence in human serum in a range of 7-100 μM, see Supplementary Table 11) and as indicated by the non-essential character of the *rpe* and *rpi* genes and the very low fitness for Tkl.

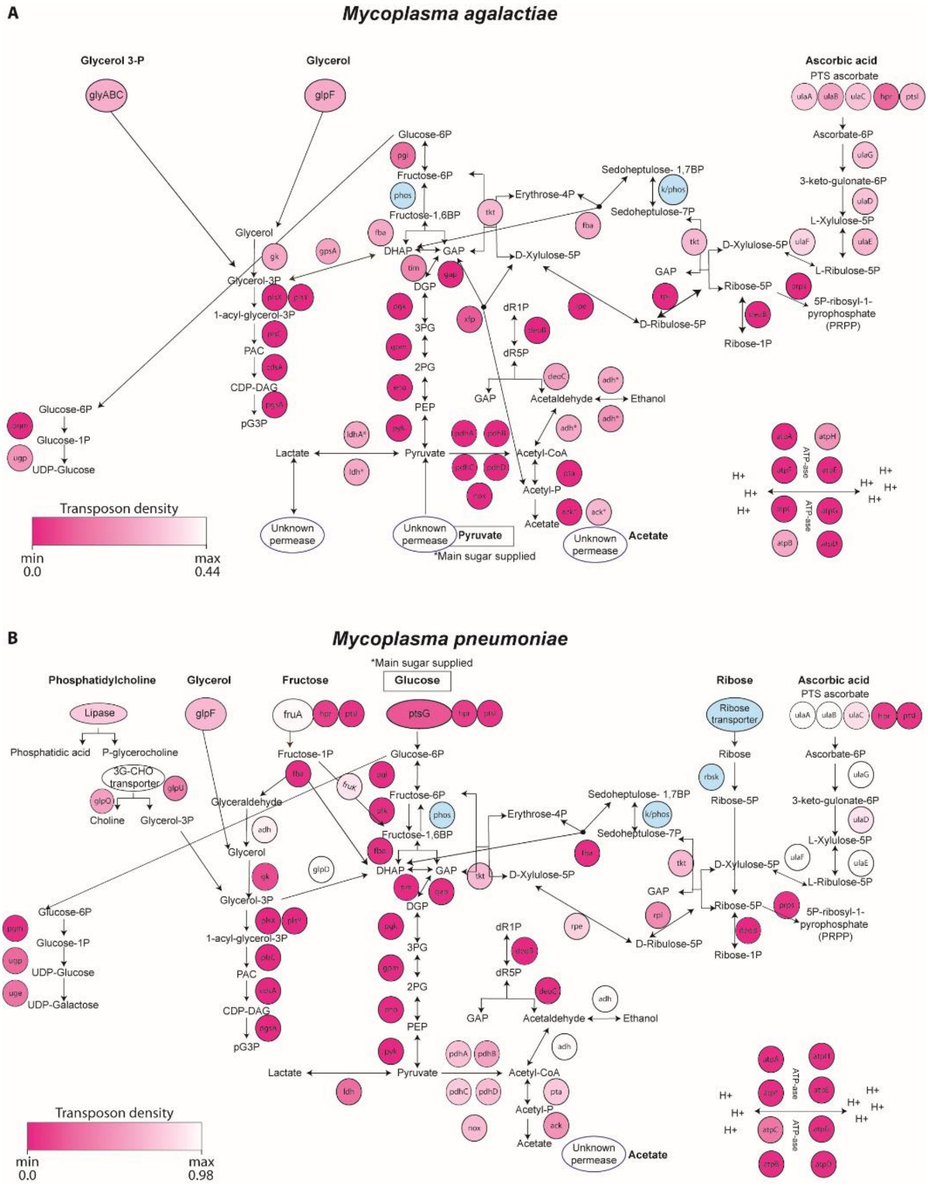


Figure 5. Insertional transposon density landscape for M. agalactiae and M. pneumoniae for selected metabolic enzymes. A, B *Sets of metabolic genes in M. agalactiae in presence of pyruvate (**A**) or in M. pneumoniae in presence of glucose (**B**) are represented in a color gradient, with gene-insertional transposon density adjusted to the specific coverage obtained for each of the strains. Minimal transposition density is indicated by magenta (showing a more essential profile) and maximum transposon density is represented in white (showing a non-essential*

profile). Blue circles indicate predicted genes but still unassigned. Duplicated genes for M. agalactiae are displayed with an asterisk (*).

The whole metabolic pathway of ascorbic acid is clearly non-essential in both species. This pathway would represent a route for generation of nucleotide precursors that would not require a connection of GAP to the PPP. This suggests that the amount of ascorbic acid present in the medium in which both species are grown is rather low (see Supplementary Table 11) and consequently, that the generation of DNA and RNA precursors is performed mainly through the link of glycolysis/gluconeogenesis to the PPP in *M. agalactiae* and/or by importing ribose in *M. pneumoniae.*

Lastly, looking into the glycerol pathway, we found that the glycerol permease (GlpF) is fitness in both *M. agalactiae* and *M. pneumoniae*. We also saw that GlpD, which links glycerol to glycolysis in *M. pneumoniae*, was not essential, while GpsA, which does the same in a reversible manner, is fitness in *M. agalactiae.* In *M. pneumoniae*, we found that glucose is not converted to glycerol-3P (Maier *et al*, 2013). Glycerol-3P is essential for generating phospholipids and therefore it should be provided by transport and phosphorylation of glycerol and/or by hydrolysis of phosphatidylcholine present in the serum that complements the medium. These two possible ways of obtaining glycerol 3-P explains the fitness character of both pathways observed in *M. pneumoniae.* In contrast, *M. agalactiae* does not have the enzymes required for production of glycerol-3P from phosphatidylcholine, and it therefore needs to import glycerol by (i) the GlpF permease, (ii) glycerol-3P through the glycerol ABC transporter, or (iii) obtain it from DHAP via GpsA. Again, this redundancy of pathways to obtain glycerol 3-P in *M. agalactiae* explains the fitness character of the genes involved in all these pathways. In this respect, it is worth noting that a small amount of glycerol is required for growth in *M. pneumoniae*, suggesting it could have another role aside from generating glycerol-3P (Yus *et al*. 2009).

In summary, our essentiality maps agree with our metabolic maps and experimental analyses of carbon sources. We observed that *M. agalactiae* could be performing a reverse glycolysis to enter into the PPP to generate nucleotide precursors. *M. pneumoniae* uses glucose to produce ATP through the glycolytic pathway but also links this pathway to the PPP to generate nucleotide precursors.

Discussion

The ascertainment of essential and active or non-active metabolic pathways is the starting point for performing metabolic engineering for different purposes, such as the development of vaccination strains, deletion of virulence factors or the enhancement of biosynthetic capacities. However, even for *Mycoplasmas*, which are bacterial cells characterized by their limited biosynthetic pathways, metabolism constitutes an intricate and interconnected network of reactions.

In this work, we integrated genomics, proteomics and gene essentiality data at high resolution for *M. agalactiae* and *M. pneumoniae* as an example of how to elucidate metabolic active pathways without experimental metabolomics determination. These predictions are validated by FBA analysis in *M. pneumoniae* and experimental information collected previously for *M. pneumoniae* (and its closely related species, *M. gallisepticum,* and in the case of *M. agalactiae,* for the closely-related species *M. bovis* (Masukagami *et al,* 2017)). These two species share 100% of all enzymes involved in metabolic pathways linked to the use of carbon sources (Supplementary Table 1).

First, we manually reconstituted the metabolic map regarding the use of carbon sources to produce energy and its connections to nucleotides, lipids, glycolipids, and glycoproteins. Besides postulating the existence of novel players to fulfil the logics of the pathway architecture, we also propose alternative functionalities for proteins. The proposal of specific glycolytic enzymes with multiple functionalities has been contemplated previously for *Mycoplasmas* (Pollack *et al*, 2002). In *M. pneumoniae*, the pyruvate kinase produces NTPs, thereby replacing a nucleotide diphosphate kinase. This cannot happen in *M. agalactiae* when using pyruvate as carbon source, which gives gluconeogenesis but not glycolysis. Thus, we propose that acetate kinase genes could be involved in the generation of NTPs. It has been described that this enzyme can also use CDP, GDP, IDP, TDP, or UDP (BRENDA:EC2.7.2.1). We then combined all experimental evidence for active metabolic fluxes in *M. agalactiae* (mainly coming from *M. bovis* (Masukagami *et al,* 2017) and in *M. pneumoniae* (Yus *et al,* 2009, Maier *et al*, 2013 Wodke *et al,* 2013) (Figures 6A and 6B, respectively). In both cases, we have experimental data regarding: *i*) carbon sources that allow growth, *ii*) metabolites identified, and *iii*) ^{13}C-glucose and/or ^{13}C-glycerol fluxes. For *M. pneumoniae*, the majority of the metabolites shown in Figure 6 were detected, indicating that the different pathways shown are active (to various extents). We have experimental evidence showing a limited flux of glucose going into the PPP or the phospholipid pathway (Maier *et al*, 2013), and a major flux going to acetate and lactate production. However, both pathways can be used, as *M. pneumoniae* can grow not only on ribose and ascorbate but also on glycerol or phosphatidylcholine. For *M. agalactiae*, experiments with ^{13}C-glucose and ^{13}C-glycerol (from *M. bovis*; Masukagami *et al,* 2017) demonstrate that there is no flux from glucose to glycolysis, and that glycerol is used not only for making phospholipids but also to enter into gluconeogenesis, thereby linking glucose-6P to the generation of glycolipids and glycoproteins. Furthermore, the growth of this bacterium with glycerol or ascorbate shows that, in the absence of pyruvate and lactate, it can carry out at least a set of the reaction involved in the canonical glycolysis. This is further supported by our finding that the introduction of the missing Pfk and PTS-glucose transporter enzymes, enables *M. agalactiae* growth using glucose as carbon source. Thus, this bacterium could perform glycolysis or gluconeogenesis, depending on the availability of carbon sources.

We have also integrated the experimental data analyzed in this manuscript from genomics, gene transposon density, and proteomics data for *M. pneumoniae* and *M. agalactiae*, to postulate active metabolic pathways for their metabolism of carbohydrates, nucleotide precursors, glycolipids, and phospholipids when grown with glucose or pyruvate, respectively, and compared that to the ones determined experimentally (Figure 6A and Figure

6B for *M. agalactiae* and *M. pneumoniae*, respectively).

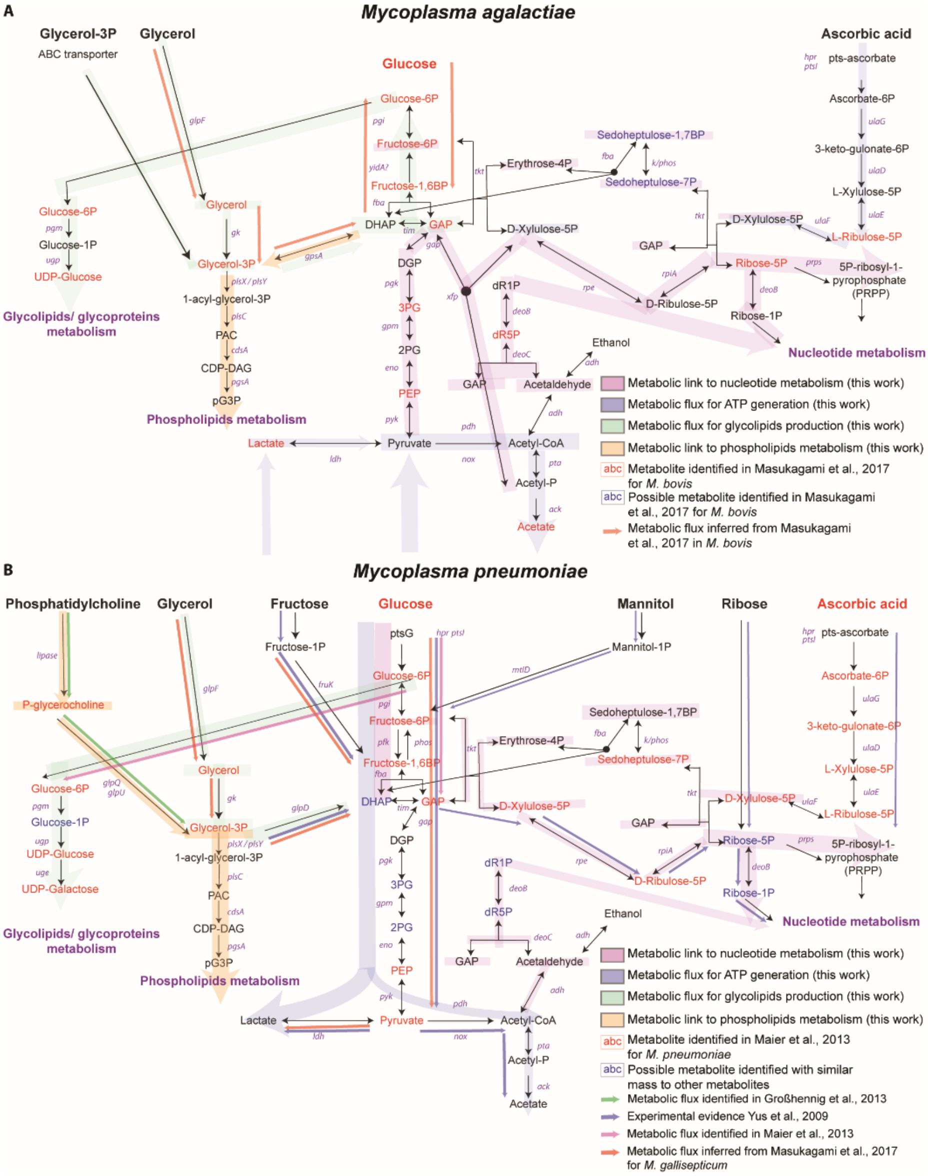


Figure 6. Integration of metabolic information collected for M. agalactiae and M. pneumoniae to date with the metabolic information observed in this work. *A For M. agalactiae, pink indicated integration of metabolic active pathways predicted in this work for nucleotide precursors; blue, metabolic pathways for obtaining energy; green, metabolic pathways for generating glycolipids or glycoproteins; and yellow, metabolic pathways linked to phospholipids metabolism. Red letters indicate metabolites detected for M. bovis, and blue letters indicate possible metabolites identified in the same work (Masukagami et al, 2017). Orange arrows indicate active*

metabolic fluxes inferred from metabolomics studies for M. bovis (Masukagami et al, 2017). B Integration of metabolic information inferred from this work for M. pneumoniae with experimental information collected in previous works. Metabolic active pathways proposed here for nucleotide metabolism are highlighted in pink, those involved in pathways for energy generation, in blue, metabolic pathways for glycolipids and glycoproteins production, in green, and phospholipids metabolism, in yellow. Red letters indicate metabolites identified previously for M. pneumoniae (Maier et al, 2013) and blue letters, metabolites whose masses make it difficult for precise ascertainment. Green arrows indicate metabolic pathways described for M. pneumoniae in Großhennig et al. (2013); pink arrows, metabolic pathways predicted in Maier et al. (2013); orange arrows, metabolic fluxes predicted for M. gallisepticum in Masukagami et al (2017); and blue arrows, active metabolic pathways ascertained from experimental validation for M. pneumoniae in Yus et al (2009).

In *M. agalactiae*, when grown on medium with pyruvate, ATP production is mediated by the conversion of pyruvate to acetate, as indicated by the essential footprint and high abundance of all enzymes involved in this metabolic conversion. The enzyme catalyzing the conversion of lactate to pyruvate (i.e. Ldh) is not essential. However, the high expression levels of Ldh in *M. agalactiae* suggest that the production of ATP from lactate might be a common phenomenon for this species. We showed that *M. agalactiae* can grow using either ascorbate or glycerol as a carbon source. However, the enzymes involved in glycerol import and its phosphorylation, or in ascorbate metabolism, have a fitness profile or are non-essential, respectively. This is logical, as pyruvate was the main carbon source in our essentiality study. Glycerol, although essential for phospholipids, can be obtained by two different transport systems as well as from gluconeogenesis.

The generation of nucleotide precursors using pyruvate as carbon source in *M. agalactiae* follows a strategy based on a gluconeogenic flux (i.e. reverse glycolysis) from pyruvate to GAP. This hypothesis is supported by the essentiality footprint found for all the genes involved this metabolite conversion, despite the fact that the glycolytic pathway is truncated in this bacterium. To finally enter into the PPP and produce in this way nucleotide precursors, GAP has to be converted to Xylulose-5P as a first step. In *M.*

agalactiae this conversion of GAP to xylulose-5P can be mediated by either Tkt or Xfp. Tkt needs fructose-6P and GAP to do this. However, we observed that although this metabolite (fructose-6P) can be produced by phosphatase hydrolysis of fructose-1,6BP, the abundance of Fba was significantly lower than that of the downstream glycolytic enzymes (Figure 3A). This, together with the low essentiality profile of Fba and the quasi-essential character of the *xfp* gene for *M. agalactiae*, seems to indicate that this conversion is mainly performed by Xfp. Indeed, this assumption is further supported by the expression levels of these enzymes in *M. agalactiae*, as that for Xfp (i.e. 29.3 log2) is much higher than that for Tkt (i.e. 24.8 log2). Thus, in M. *agalactiae*, there is a flux starting at pyruvate and ending in ribose-5P that enables the synthesis of nucleotide precursors, such as PRPP.

We postulate that synthesis of glycolipids and glycoproteins might be mediated by two different fluxes in *M. agalactiae*. We propose that in *M. agalactiae*, glucose-6P, the major building block for glycolipids synthesis, can be generated thanks to the existence of a complete gluconeogenic flux from pyruvate to glucose-6P. In this way, the consecutive activity of the glycolytic enzymes Pyk, Eno, Gpm, Pgk, Gap, and Fba, as well as a putative phosphatase, most likely the YidA-related phosphatase encoded by MAGA7784_RS03930 (Figure 6) would enable this conversion. All the genes coding for these enzymes show an essential profile (e.g. *pyk*, *eno*, *gpm*, *pgk*, and *gap*) or a quasi-essential profile (e.g. *pgi* and *fba*). Alternatively, the generation of glucose-6P could be mediated by a partial gluconeogenic flux starting at glycerol. However, as discussed in the previous section, *M. agalactiae* can only produce glycerol-3P from DHAP, or import glycerol or glycerol-3P, but it cannot produce it from phosphatidylcholine. The fitness profile of the respective enzymes suggests that these pathways are all active. Finally, we observed that all the enzymes going from glycerol-3P to phospholipids have an essential profile in both *M. agalactiae* and *M. pneumonia*e. This indicates the importance of generating glycerol-3P.

Whereas most of our predictions are based on the essentiality study done for *M. agalactiae*, the proteomics data strongly appear to corroborate our hypothesis. In fact, it is curious to observe that expression levels of all the enzymes involved in the lower part of the glycolytic pathway (i.e. Gap, Pgk, Gpm, Eno, and Pyk) are higher than the ones observed for the enzymes found in the upper part (i.e. Fba and Pgi). This observation falls in line with the fact that the lower part of the glycolytic pathway feeds both nucleotide metabolism and glycolipids metabolism, whereas the enzymes in the upper part are intended to exclusively feed glycolipid metabolism. In contrast, *M. pneumoniae* shows a different metabolic strategy. In this species, ATP generation is mediated through glycolysis, as indicated by the protein abundance and having an essential footprint present of all the enzymes in the whole pathway from glucose to pyruvate. Through this set of reactions, two molecules of ATP are obtained, which can be increased to four if pyruvate is further metabolized to acetate rather than to lactate. Alternatively, under glucose depletion conditions, *M. pneumoniae* can also use alternative carbon sources, such as fructose or mannose, which enter glycolysis at different point-load levels.

Regarding the PPP, *M. agalactiae* can link glycolysis and the PPP using the activity of either Tkt or Xfp, but the genome of *M. pneumoniae* only encodes for Tkt. Intriguingly, however, the *tkt* gene shows a low fitness character in both species, which makes sense for *M. agalactiae*, given the alternative route based on Xfp, but not for *M. pneumoniae*, which lacks the *xfp* gene. The reason could be that *M. pneumoniae* can use ribose as carbon source as well as import nucleosides (Yus *et al*, 2009), while *M. agalactiae* does not grow with ribose. Finally, the link to glycolipids metabolism is quite obvious in *M. pneumoniae* given its ability to directly import and phosphorylate glucose into glucose-6P which can be later employed for glycolipids synthesis. Regarding phospholipids in *M. pneumoniae* the enzymes downstream of glycerol-3P are essential and have as *M. agalactiae* similar expression levels, while those involved in glycerol-3P generation have a fitness profile, as the bacterium has three possible routes: from fructose-1P, from phosphatidylcholine, and by importing it directly. Essentiality analysis suggests that the fructose flux must be very low. In this case, glycolysis cannot produce glycerol-3P, as the GlpD enzyme produces DHAP from glycerol-3P in an irreversible manner.

The study of bacterial metabolism may give a hint about their lifestyle within a host. In the case of *M. pneumoniae,* a lung-colonizing bacterium, the essentiality profile for its active metabolic pathways might be different in *in vivo* conditions, where phosphatidylcholine and phosphatidylglycerol are the main carbon sources present in the lung alveolar surfactant. The fact that *M. agalactiae* or *M. bovis* have the ability to metabolize acetate or pyruvate, products of glycolysis could suggest that they could internalize into the cytoplasm of eukaryotic cells (Virulence, persistence and dissemination of Mycoplasma bovis, 2015) and use pyruvate as substrate.

Summing up, integration of the metabolic pathway genomics, the experimental analysis on carbon sources supporting growth, the quantitative proteome and the essentiality analyses allows us to predict the metabolic pathways active under the conditions studied. Notably, our predictions agree not only with the metabolic studies done in *M. pneumoniae* and *M. bovis,* but also with the experimental validation of carbon sources capable to promote bacterial growth done in this study. Protein quantification together with gene essentiality under different experimental conditions could thus be a way to complement metabolic flux analyses that use ^{13}C-tagged compounds or metabolic modeling. To conclude, it should be noted that although in some aspects the information provided by proteomics and gene essentiality might seem redundant, both approaches support each other and offer complementary information. Gene essentiality shows essential pathways and when more than one pathway is connected to an essential metabolite indicates if they can replace each other. Protein abundance in pathways connected to a metabolite gives an indication of the preferred flow. Combination of both could offer new insights. In *M. pneumoniae* the expression levels of proteins converting pyruvate into lactate and acetate are comparable. Notwithstanding, the essentiality profile of those genes differs, showing *ldh* a lower density of transposon insertions and therefore pointing to lactate as main metabolism sub-product for *M. pneumoniae.* On the other hand, there are other situations in which proteomics data could clarify the information provided by gene essentiality. For instance, when there is redundancy in the pathways to produce a given metabolite such as xylulose 5-P in *M. agalactiae*, the essentiality profile could suggest which pathway is absorbing most of the flux (i.e. *xfp* is much more essential than the alternative pathway based on *fba* and *tkt* encoded activities), but this can be further corroborated with protein abundance (i.e. Xfp is detected at levels almost ten times higher than Tkt).

Materials and Methods

Bacteria strains and general culture conditions. The strain 7784 was used for wild-type *M. agalactiae* 7784 (kindly provided by Dr. Christine Citti), and the strain M129, for wild-type *M. pneumoniae (*ATTC 29342, subtype 1). Both species were grown at 37ºC in standard Hayflick medium (for each adjusted 500 ml of Hayflick medium, at pH 7.6 it contains 7.3 g PPLO broth, 10 μg phenol red, 11.9 g HEPES, 100 ml inactivated fetal horse serum, and 500,000 units sterile penicillin G) supplemented with 0.5% sodium pyruvate, pH 7.6 (Sigma-Aldrich) for *M. agalactiae* or 0.5% glucose, pH 7.6 (Sigma-Aldrich) for *M. pneumoniae*. When indicated, *M. agalactiae* was grown in ‘depleted Hayflick medium’ supplemented with the desired carbon source. *M. agalactiae* was grown in suspension (180 rpm, 37ºC), while *M. pneumoniae* was grown attached to a flask without shaking at 37ºC.

Experimental metabolic analysis of *M. agalactiae*. For all experiments, an inoculum of 2 µg *M. agalactiae* cells was used per ml of medium. “Depleted medium” was generated by inoculating 2 L of complete Hayflick medium (lacking pyruvate) with *M. agalactiae*, growing at 37 ºC with 180 rpm orbital shaking for 48 h, centrifuging twice at 9408 g for 10 min at 4 ºC, discarding the pellet, and filtering the medium using Stericup®-GP Filter Units polyethersulfone membrane (0.22 µm pore size) (Millipore Express® PLUS, Z660507-12EA). Depleted medium was stored at 4ºC until use, and a single batch was used for all experiments in this work when indicated. Different carbon sources were prepared in concentrated stocks and then diluted to a fixed final concentration. All components analyzed were dissolved in sterile Milli-Q water except L-α-phosphatidylcholine, which was dissolved in ethanol (Merck, 108543). The final concentration used was 0.174% for L-arginine (Sigma-Aldrich, A5006-100), 4x10^{-3}% for L-α-phosphatidylcholine (Sigma-Aldrich, 61755-25G), 0.05% for glycerol (Sigma-Aldrich, G5516-1L), and 0.1% for D-fructose (Sigma-Aldrich, F0127-10MG), D-glucose (Sigma-Aldrich, G8270-1KG), D-ribose (Sigma-Aldrich, R1757-10G-A), D-mannose (Sigma-Aldrich, M6020-25G), D-mannitol (Sigma-Aldrich, M4125-100G), sodium pyruvate (pH adjusted to 7.8; Sigma-Aldrich, P2256-25G), L-ascorbic acid (pH adjusted to 7.8; Sigma-Aldrich, A0278-25G), and L-lactic acid (provided in solution, pH adjusted to 7.3; Sigma-Aldrich, 199257-5G,). For each culture, *M. agalactiae* stock was inoculated in 50 ml Falcon tubes (Fisher Scientific, 14-432-22) filled with 10 ml depleted medium supplemented with the different carbon sources. After 48 h of growth, cultures were centrifuged twice and washed with one ml PBS before final resuspension in 250 µl of lysis solution (4% SDS, 100 mM HEPES). Growth was measured by protein content using the Pierce™ BCA Protein Assay Kit (Thermo Fisher Scientific, 23225) following manufacturer's protocol. For growth curves based on protein content for *M. agalactiae WT* and growth comparison of *M. agalactiae* WT in Hayflick medium versus the depleted version see Supplementary Table 3.

Generation of a *M. agalactiae* strain able to metabolize glucose. Wild-type *M. agalactiae* was transformed via transposon with two *M. pneumoniae* genes using plasmids described in Supplementary Table 5. The transposon plasmid was generated following the Gibson assembly method (Gibson *et al*, 2009). DNA was isolated from NEB® 5-alpha High Efficiency (C2987P). Clones were isolated using LB agar plus ampicillin (100 µg/ml) plates and confirmed by sequencing (Eurofins Genomics). Cultures of *M. agalactiae* (25 ml) were grown 48 h in Hayflick medium. After, medium was removed, and 10 ml fresh Hayflick medium was added. After 3 h, cultures were centrifuged at 9408 g at 4ºC, washed three times with chilled electroporation buffer (EB; 272 mM sucrose, 8 mM HEPES, pH 7.4), and resuspended in 300 µl chilled EB. Cells (50 µl) were then mixed with 1.5 µg DNA, incubated 20 min on ice, and electroporated in 0.1-cm electro cuvettes using a BIO-RAD Gene Pulser Xcell apparatus, using the settings of 1250 V / 25 µF / 100 Ω. Immediately after the pulse, 420 µl fresh Hayflick was added, and cells were incubated for 90 min (*M. agalactiae*) at 37ºC before seeding on agar plates. Culture (100 µl) was then inoculated into 10 ml Hayflick medium with 100 µg/ml gentamicin, 0.25% glucose, and 0.25% pyruvate for a progressive metabolic adaptation. After 4 days, cultures were centrifuged at 9408 g, medium was discarded, and the cell pellet was resuspended in 10 ml Hayflick medium with 1% glucose and 100 µg/ml gentamicin and grow further. After 2 days, the culture of *M. agalactiae* strain was collected in 1 ml of Hayflick medium to generate the primary stock of cells. The expression of both heterologous proteins from *M. pneumoniae* in *M. agalactiae* have been detected by mass spectrometry (Supplementary Table 6). The growth of *M. agalactiae* synthetic strain double mutant (expressing both heterologous proteins) or single mutants (expressing only single protein; or MPN207 or MPN302) has been assessed by protein content after 24 hours of growth in the depleted version of the medium supplemented with glucose (Supplementary Table 3).

Secondary metabolite measurement. For each metabolite the concentration was measured at the initial point and then after 48 h of growth. *M. agalactiae* and *M. pneumoniae* cultures were grown for 48 h in standard Hayflick medium (supplemented with pyruvate or glucose, respectively). For *M. agalactiae,* the culture was centrifuged at 4ºC for 10 min at 9408 g, and the supernatant was collected and placed on ice. For *M. pneumoniae* (which grows attached), supernatant was collected directly and placed on ice. Concentrations were measured for ethanol with the enzymatic kit for ethanol assay (Megazyme, K-ETOH), for glucose with the Glucose Colorimetric/Fluorometric Assay Kit (Biovision, #K606), for L-lactate with the Lactate Colorimetric/Fluorometric Assay Kit (Biovision, #K607), and for acetate with the Acetate Colorimetric Assay Kit (Biovision, #K658); all methods followed the manufacturers' instructions. Analyses of the different metabolites were performed on different days and collected in Supplementary Table 4.

Preparing samples for mass spectrometry analysis. For *M. pneumoniae,* 50 µg of cells were inoculated in two T25 flask with 5 ml Hayflick medium supplemented with glucose 0.5% and grown for 48 h. Cells were then washed twice with PBS and scraped for final collection in a 1.5 ml collection tube. For *M. agalactiae*, 5 ml cultures inoculated in Falcon tubes were grown for 48 h in standard Hayflick medium supplemented with sodium pyruvate 0.5%. Cultures were then centrifuged and washed twice in PBS. For both species, pelleted cells were then resuspended ten times with a 25-gauge needle to separate individual cells and resuspended in 150 µl of urea lysis buffer (6 M urea, 0.2 M NH_4CO_3) or SDS lysis buffer (4% SDS, 100 mM HEPES). For samples resuspended in urea buffer, two biological replicates were performed for each species. The protein concentration of each lysate was measured using the Pierce BCA Protein kit following manufacturer's protocol. For urea samples, 10 µg samples were reduced with 30 nmol dithiothreitol (37 ºC, 60 min) and alkylated in the dark with 60 nmol iodoacetamide (25ºC, 30 min). The resulting protein extract was first diluted to 2 M urea with 200 mM ammonium bicarbonate for digestion with endoproteinase LysC (1:10 w:w, overnight at 37ºC; Wako, cat # 129-02541) and then diluted 2-fold with 200 mM ammonium bicarbonate for trypsin digestion (1:10 w:w, 8 h at 37ºC; Promega, cat # V5113). Samples with SDS lysis buffer were reduced with 90 nmol dithiothreitol (30 min at 56°C) and alkylated in the dark with 180 nmol iodoacetamide (30 min at 25°C). Following the filter-aided sample preparation (FASP) method described in (Kwasniewski 2016), samples were then digested with 3 µg LysC (Wako, cat # 129-02541) overnight at 37°C, and then with 3 µg of trypsin (Promega, cat # V5113) for eight hours at 37°C. After digestion, peptide mixes were acidified with formic acid and desalted with a MicroSpin C18 column (The Nest Group, Inc) prior to LC-MS/MS analysis.

Chromatographic and mass spectrometric analyses. Samples were analyzed using a LTQ-Orbitrap Velos Pro-mass spectrometer (Thermo Fisher Scientific, San Jose, CA, USA) coupled to an EASY-nLC 1000 (Thermo Fisher Scientific (Proxeon), Odense, Denmark). Peptides were loaded onto a 2-cm Nano Trap column with an inner diameter of 100 µm packed with ^{18}C particles of 5 µm particle size (Thermo Fisher Scientific) and separated by reversed-phase chromatography using a 25-cm column with an inner diameter of 75 µm, packed with 1.9 µm ^{18}C particles (Nikkyo Technos Co., Ltd. Japan). Chromatographic gradients started at 93% buffer A, 7% buffer B with a flow rate of 250 nl/min for 5 min, and gradually increased to 65% buffer A, 35% buffer B over 120 min. After each analysis, the column was washed for 15 min with 10% buffer A, 90% buffer B. Buffer A was 0.1% formic acid in water, and buffer B, 0.1% formic acid in acetonitrile. The mass spectrometer was operated in positive ionization mode with nano spray voltage set at 2.1 kV and source temperature at 300°C. Ultramark 1621 was used for external calibration of the FT mass analyzer prior to analyses, and an internal calibration was performed using the background polysiloxane ion signal at m/z 445.1200. Acquisition was performed in data-dependent acquisition (DDA) mode, and full MS scans with 1-micro scans at a 60,000 resolution were used over a mass range of m/z 350–2000 with detection in the Orbitrap. Auto gain control (AGC) was set to 1 x10^6, dynamic exclusion (60 seconds) and charge state filtering disqualifying singly charged peptides was activated. In each cycle of DDA analysis, following each survey scan, the top twenty most intense ions with multiple charged ions above a threshold ion count of 5000 were selected for fragmentation. Fragment ion spectra were produced via collision-induced dissociation (CID) at normalized collision energy of 35% and they were acquired in the ion trap mass analyzer. AGC was set to 1x 10^4, and an isolation window of 2.0 m/z, an activation time of 10 ms, and a maximum injection time of 100 ms were used. Data was acquired with Xcalibur software v2.2. Digested bovine serum albumin (New England Biolabs cat # P8108S) and was analyzed between each sample run to avoid sample carryover and to assure stability of the instrument. Finally QCloud (Chiva *et al*, 2018) was used to control instrument longitudinal performance during the project.

Data analysis. Acquired spectra were analyzed using the Proteome Discoverer software suite (v2.0, Thermo Fisher Scientific) and the Mascot search engine (v2.6, Matrix Science (Perkins *et al*, 1999)). Data were searched against a *M. agalactiae* strain 7784 database (76,678 entries) or *M. pneumoniae* M129 database (87,070 entries) plus a list of common contaminants and all the corresponding decoy entries (Perkins *et al*, 1999; Beer *et al*, 2017). For peptide identification, a precursor ion mass tolerance of 7 ppm was used for the MS1 level, trypsin was used as an enzyme, and up to three missed cleavage sites were allowed. The fragment ion mass tolerance was set to 0.5 Da for MS2 spectra. Oxidation of methionine and N-terminal protein acetylation were used as variable modifications, with carbamidomethylation on cysteine set as a fixed modification. False discovery rate (FDR) in peptide identification was set to a maximum of 5%. Peptide quantification data were retrieved from the "Precursor ion area detector" node from Proteome Discoverer (v2.0) using 2 ppm mass tolerance for the peptide extracted ion current (XIC). The obtained values were used to calculate protein top 3 area with the unique peptide for ungrouped proteins. The raw proteomics data were deposited in the PRIDE repository (Vizcaíno *et al*, 2016) with the dataset identifier PXD015800. The obtained dataset was normalized assuming equal total protein content for both species (using two biological replicates per species). Total protein content of sample 170718_TFLS_01_01 from *M. pneumoniae* was used as a reference for normalization (Supplementary Table 1 and 7). The average for the area obtained for each of the protein have been calculated and log2 values were then obtained for both biological replicates. The fold change in log2 protein abundance have been calculated for each of the species. The t-test statistical analysis have been performed for each of the selected orthologous genes, two-tailed distribution and assuming unequal variances; later we corrected the obtained p-values for multiple tests using Benjamini-Hochberg correction. For the specific comparison of individual orthologous genes p-values were calculated assuming one-tailed distribution and assuming unequal variances. Orthology between *M. pneumoniae* M129, *M. agalactiae* 7784 and *M. bovis* PG45 was generated using the Microbial Genomes Database (MBGD) server considering default parameters and curated for duplicated using BlastP. We added up mass spectrometry areas of duplicated forms in *M. agalactiae* when the two copies were conserved respect the same protein in *M. pneumoniae* (Supplementary Table 1). The average area under the curve of the three best-flying peptides was used to estimate the relative concentration of each protein. Data for both species were normalized assuming equal total protein content. As proteins highlighted with asterisk (*) in Figure 3 were not detected in the urea samples, the values obtained in SDS are shown after normalization, with an assumption of equal total protein content. The protein encoded by gene MPN637 in *M. pneumoniae* was not detected in these experiments but data was inferred from other *M. pneumoniae* proteomics datasets published in (Miravet-Verde *et al*, 2019) with an average value of 25.75 log2. To represent protein abundance in Figure 3, all enzymes with an expression value of lower than 27 log2 were arbitrarily represented with same color code (green) to indicate "minimum". Note that below this value, peptide detection becomes random, and protein concentrations cannot be accurately estimated. Thus, in Figure 3, protein expression levels range from 27 to 32.6 log2. The upper value corresponds to the maximum expression value obtained for the set of genes involved in sugar metabolism for the two species and is represented in magenta color (i.e. *ldh,* MAGA7784_RS02675).

Protein abundance variability analysis. In order to study if the protein abundance variability is reduced when considering proteins pertaining to a unique linear metabolic pathway, we extracted eight linear sub-pathways from the studied set of metabolic pathways (Supplementary Table 1). Please note that for *M. agalactiae* the number of sub-pathways analyzed is seven because enzyme Pgi was included in linear sub-pathway one (genes involved in glycolipids) while in *M. pneumoniae* Pgi was grouped with Pfk, enzyme absent in *M. agalactiae*.

The selection was based on single entry and single exit genes (nodes with one entry edge and one exit edge) (Supplementary Table 8). For both species separately, the coefficient of variation was calculated (CV, standard deviation corrected by the mean) and a background model of expected CVs by random sampling (1000 iterations with each of them a size equal to the number of proteins in the sub-pathway analyzed). The same process was repeated using raw quantification (natural) and using a $\log_2$ transformation (log) (Supplementary Figure 2A and 2B for *M. pneumoniae* and Supplementary Figure 2C and 2D for *M. agalactiae*). In parallel, we merged all sub-pathways together (independent for each species and alternatively with raw values and logarithmic values) and compared using Wilcoxon signed-rank tests the distribution of observed CVs to a distribution of average expected CVs by each linear sub-pathway (Supplementary Table 8, Figure 4). P-values obtained for *M. pneumoniae* for raw and log2 values are p-value=0.005 and p-value=0.012, respectively. P-values obtained for *M. agalactiae* for raw and log2 values are p-value=0.045 and p-value=0.008, respectively

To unravel whether the similarity in protein abundances observed for enzymes involved in the same sub-pathway arises as a consequence of belonging the same transcriptional unit we showed operon distributions in *M. pneumoniae* from previously available literature (Güell *et al*, 2009 and Junier *et al*, 2016). In the case of *M. agalactiae*, operon distribution information was collected from DOOR database, which predicts operon organization based on genomic features (Mao *et al,* 2009) using as reference organism the strain NC_013948.

The information of operon distribution for both strains is available in Supplementary Table 9.

Gene essentiality. Transposon density values were different for both species, with a higher coverage obtained in the *M. pneumoniae* (Supplementary Table 10 for the reanalyzed *M. pneumoniae* dataset) compared to *M. agalactiae* dataset (Montero-Blay *et al*, 2019). In both datasets, the culture medium used to perform the experiments was Hayflick supplemented with 0.5% glucose in the case of *M. pneumoniae* or 0.5% sodium pyruvate for *M. agalactiae*. The maximum values of transposon density observed for a gene in these species were 0.98 and 0.40 for *M. pneumoniae* and *M. agalactiae*, respectively. To represent these values in Figure 4, the range of transposon density values was normalized for each species independently. In Figure 4, colors are shown along a min–max scale after normalization for each of the species (from low transposon density in magenta, to high transposon density in white).

Acknowledgements

The proteomics analyses were performed in the CRG/UPF Proteomics Unit, which is part of the Proteored, PRB3, and is supported by grant PT 17/0019, of the PE I+D+i 2013-2016, funded by ISCIII and ERDF. We are especially grateful to Guadalupe Espadas-Garcia and Eduard Sabidó for technical and scientific guidance for proteomics. We thank V.A. Raker for manuscript editing. This project was financed by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program, under grant agreement Nos 634942 (MycoSynVac) and 670216 (MYCOCHASSIS).

Author Contributions

AMB performed the experiments. AMB and SMV performed data analysis. LS, MLS and CPL contributed with ideas and direct supervision of the project. All authors participated in evaluating results and discussions about the project. AMB, CPL and LS wrote the paper. AMB and SMV prepared tables. AMB prepared figures. LS, CPL, SMV and MLS reviewed the manuscript. All authors read and approved the final manuscript.

References

Beale DJ, Pinu FR, Kouremenos KA, Poojary MM, Narayana VK, Boughton BA, Kanojia K, Dayalan S, Jones OAH & Dias DA (2018) Review of recent developments in GC–MS approaches to metabolomics-based research. *Metabolomics* **14:** Available at: http://dx.doi.org/10.1007/s11306-018-1449-2

Beer LA, Liu P, Ky B, Barnhart KT & Speicher DW (2017) Efficient Quantitative Comparisons of Plasma Proteomes Using Label-Free Analysis with MaxQuant. *Methods Mol. Biol.* **1619:** 339–352

Chiva C, Olivella R, Borràs E, Espadas G, Pastor O, Solé A & Sabidó E (2018) QCloud: A cloud-based quality control system for mass spectrometry-based proteomics laboratories. *PLOS ONE* **13:** e0189209 Available at: http://dx.doi.org/10.1371/journal.pone.0189209

Gibson DG, Young L, Chuang R-Y, Venter JC, Hutchison CA 3rd & Smith HO (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6:** 343–345

Glass JI, Assad-Garcia N, Alperovich N, Yooseph S, Lewis MR, Maruf M, Hutchison CA 3rd, Smith HO & Venter JC (2006) Essential genes of a minimal bacterium. *Proc. Natl. Acad. Sci. U. S. A.* **103:** 425–430

Großhennig S, Schmidl SR, Schmeisky G, Busse J & Stülke J (2013) Implication of glycerol and phospholipid transporters in Mycoplasma pneumoniae growth and virulence. *Infect. Immun.* **81:** 896–904

Güell M, van Noort V, Yus E, Chen W-H, Leigh-Bell J, Michalodimitrakis K, Yamada T, Arumugam M, Doerks T, Kühner S, Rode M, Suyama M, Schmidt S, Gavin A-C, Bork P & Serrano L (2009) Transcriptome complexity in a genome-reduced bacterium. Science **326:** 1268–1271

Himmelreich R, Hilbert H, Plagens H, Pirkl E, Li BC & Herrmann R (1996) Complete sequence analysis of the genome of the bacterium Mycoplasma pneumoniae. Nucleic Acids Res. **24**: 4420–4449

Hsu TM, Welner DH, Russ ZN, Cervantes B, Prathuri RL, Adams PD & Dueber JE (2018) Employing a biochemical protecting group for a sustainable indigo dyeing strategy. *Nature Chemical Biology* **14:** 256–261 Available at:

http://dx.doi.org/10.1038/nchembio.2552

Jang C, Chen L & Rabinowitz JD (2018) Metabolomics and Isotope Tracing. *Cell* **173:** 822–837

Junier I, Unal EB, Yus E, Lloréns-Rico V & Serrano L (2016) Insights into the Mechanisms of Basal Coordination of Transcription Using a Genome-Reduced Bacterium. Cell Syst **2**: 391–401

Kamminga T, Slagman S-J, Bijlsma JJE, Martins Dos Santos VAP, Suarez-Diez M & Schaap PJ (2017) Metabolic modeling of energy balances in Mycoplasma hyopneumoniae shows that pyruvate addition increases growth rate. Biotechnol. Bioeng. **114**: 2339–2347

Kanehisa M & Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. **28**: 27–30

Karp PD, Weaver D & Latendresse M (2018) How accurate is automated gap filling of metabolic models? BMC Syst. Biol. **12**: 73

Kühner S, van Noort V, Betts MJ, Leo-Macias A, Batisse C, Rode M, Yamada T, Maier T, Bader S, Beltran-Alvarez P, Castaño-Diez D, Chen W-H, Devos D, Güell M, Norambuena T, Racke I, Rybin V, Schmidt A, Yus E, Aebersold R, et al (2009) Proteome Organization in a Genome-Reduced Bacterium. *Science* **326:** 1235–1240 Available at: http://dx.doi.org/10.1126/science.1176343

Kuznetsova E, Proudfoot M, Gonzalez CF, Brown G, Omelchenko MV, Borozan I, Carmel L, Wolf YI, Mori H, Savchenko AV, Arrowsmith CH, Koonin EV, Edwards AM & Yakunin AF (2006) Genome-wide analysis of substrate specificities of the Escherichia coli haloacid dehalogenase-like phosphatase family. *J. Biol. Chem.* **281:** 36149–36161

Lluch-Senar M, Delgado J, Chen W-H, Lloréns-Rico V, O'Reilly FJ, Wodke JA, Unal EB, Yus E, Martínez S, Nichols RJ, Ferrar T, Vivancos A, Schmeisky A, Stülke J, van Noort V, Gavin A-C, Bork P & Serrano L (2015) Defining a minimal cell: essentiality of small ORFs and ncRNAs in a genome-reduced bacterium. *Mol. Syst.*

Biol. **11:** 780

Long CP & Antoniewicz MR (2019) High-resolution 13C metabolic flux analysis. *Nature Protocols* 14: 2856–2877 Available at: http://dx.doi.org/10.1038/s41596-019-0204-0

Maier T, Marcos J, Wodke JAH, Paetzold B, Liebeke M, Gutiérrez-Gallego R & Serrano L (2013) Large-scale metabolome analysis and quantitative integration with genomics and proteomics data in Mycoplasma pneumoniae. *Mol. Biosyst.* **9:** 1743–1755

Mao F, Dam P, Chou J, Olman V & Xu Y (2009) DOOR: a database for prokaryotic operons. Nucleic Acids Res. **37**: D459–63

Masukagami Y, De Souza DP, Dayalan S, Bowen C, O'Callaghan S, Kouremenos K, Nijagal B, Tull D, Tivendale KA, Markham PF, McConville MJ, Browning GF & Sansom FM (2017a) Comparative Metabolomics of Mycoplasma bovis and Mycoplasma gallisepticum Reveals Fundamental Differences in Active Metabolic Pathways and Suggests Novel Gene Annotations. *mSystems* **2:** Available at: http://dx.doi.org/10.1128/msystems.00055-17

Masukagami Y, De Souza DP, Dayalan S, Bowen C, O'Callaghan S, Kouremenos K, Nijagal B, Tull D, Tivendale KA, Markham PF, McConville MJ, Browning GF & Sansom FM (2017b) Comparative Metabolomics of and Reveals Fundamental Differences in Active Metabolic Pathways and Suggests Novel Gene Annotations. *mSystems* **2:** Available at: http://dx.doi.org/10.1128/mSystems.00055-17

Miravet-Verde S, Ferrar T, Espadas-García G, Mazzolini R, Gharrab A, Sabido E, Serrano L & Lluch-Senar M (2019) Unraveling the hidden universe of small proteins in bacterial genomes. *Mol. Syst. Biol*. 15: e8290

Montero-Blay A, Miravet-Verde S, Lluch-Senar M, Piñero-Lambea C & Serrano L (2019) SynMyco transposon: engineering transposon vectors for efficient transformation of minimal genomes. *DNA Research* Available at: http://dx.doi.org/10.1093/dnares/dsz012

Ochsner AM, Christen M, Hemmerle L, Peyraud R, Christen B & Vorholt JA (2017) Transposon Sequencing Uncovers an Essential Regulatory Function of Phosphoribulokinase for Methylotrophy. *Curr. Biol.* **27:** 2579–2588.e6

Perkins DN, Pappin DJ, Creasy DM & Cottrell JS (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20:** 3551–3567

Pettersson B, Uhlén M & Johansson KE (1996) Phylogeny of some mycoplasmas from ruminants based on 16S rRNA sequences and definition of a new cluster within the hominis group. *Int. J. Syst. Bacteriol.* **46:** 1093–1098

Pollack JD, Myers MA, Dandekar T & Herrmann R (2002) Suspected utility of enzymes with multiple activities in the small genome Mycoplasma species: the replacement of the missing 'household' nucleoside diphosphate kinase gene and activity by glycolytic kinases. *OMICS* **6:** 247–258

Ponce-de-León M, Apaolaza I, Valencia A & Planes FJ (2019) On the inconsistent treatment of gene-protein-reaction rules in context-specific metabolic models. Bioinformatics Available at: http://dx.doi.org/10.1093/bioinformatics/btz832

Qi J, Guo A, Cui P, Chen Y, Mustafa R, Ba X, Hu C, Bai Z, Chen X, Shi L & Chen H (2012) Comparative geno-plasticity analysis of Mycoplasma bovis HB0801 (Chinese isolate). *PLoS One* **7:** e38239

Rechnitzer H, Rottem S & Herrmann R (2013) Reconstitution of an active arginine deiminase pathway in Mycoplasma pneumoniae M129. *Infect. Immun.* **81:** 3742–3749

Rottem S (2003) Interaction of Mycoplasmas With Host Cells. *Physiological Reviews* **83:** 417–432 Available at: http://dx.doi.org/10.1152/physrev.00030.2002

Virulence, persistence and dissemination of Mycoplasma bovis (2015) Vet. Microbiol. **179**: 15–22

Vanyushkina AA, Fisunov GY, Gorbachev AY, Kamashev DE & Govorun VM (2014) Metabolomic analysis of three Mollicute species. PLoS One **9**: e89312

Vizcaíno JA, Csordas A, Del-Toro N, Dianes JA, Griss J, Lavidas I, Mayer G, Perez-Riverol Y, Reisinger F, Ternent T, Xu Q-W, Wang R & Hermjakob H (2016) 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **44:** 11033

Wodke JAH, Puchałka J, Lluch-Senar M, Marcos J, Yus E, Godinho M, Gutiérrez-Gallego R, dos Santos VAPM, Serrano L, Klipp E & Maier T (2013) Dissecting the energy metabolism in Mycoplasma pneumoniae through genome-scale metabolic modeling. *Mol. Syst. Biol.* **9:** 653

Yang H, Krumholz EW, Brutinel ED, Palani NP, Sadowsky MJ, Odlyzko AM, Gralnick JA & Libourel IGL (2014) Genome-scale metabolic network validation of Shewanella oneidensis using transposon insertion frequency analysis. *PLoS Comput. Biol.* **10:** e1003848

Yus E, Maier T, Michalodimitrakis K, van Noort V, Yamada T, -H. Chen W, Wodke JAH, Guell M, Martinez S, Bourgeois R, Kuhner S, Raineri E, Letunic I, Kalinina OV, Rode M, Herrmann R, Gutierrez-Gallego R, Russell RB, -C. Gavin A, Bork P, et al (2009) Impact of Genome Reduction on Bacterial Metabolism and Its Regulation. *Science* **326:** 1263–1268

Zimmermann M, Zimmermann-Kogadeeva M, Wegmann R & Goodman AL (2019) Separating host and microbiome contributions to drug pharmacokinetics and toxicity. *Science* **363:** Available at: http://dx.doi.org/10.1126/science.aat9931

To access all the material available for this article, please visit the following QR.

4. Chapter 4: Towards the use of *M. pneumoniae* as immunomodulatory agent in the lung

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]		
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	
[redacted]	[redacted]	
[redacted]	[redacted]	[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

	[redacted]		[redacted]		[redacted]
	[redacted]		[redacted]		[redacted]
	[redacted]		[redacted]		[redacted]
	[redacted]		[redacted]		[redacted]
	[redacted]		[redacted]		[redacted]
	[redacted]		[redacted]		[redacted]
	[redacted]		[redacted]		[redacted]
	[redacted]		[redacted]		[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]

[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]

[redacted]

[redacted] Chapter 5. [redacted]

The work showed in this chapter has been performed with the collaboration of Dr. Javier Delgado, ModelX creator and developer, and Prof. Luis Serrano, FoldX creator, and developer.

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]

[redacted]

[redacted]

[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]			
[redacted]	[redacted]	[redacted]			
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]			
[redacted]	[redacted]	[redacted]			
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]			
[redacted]	[redacted]	[redacted]			

[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

[redacted]

[redacted]

[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

[redacted]

[redacted]

[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]

[redacted]

[redacted]

[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

6. Chapter 6. [redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

[redacted]

[redacted]

[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

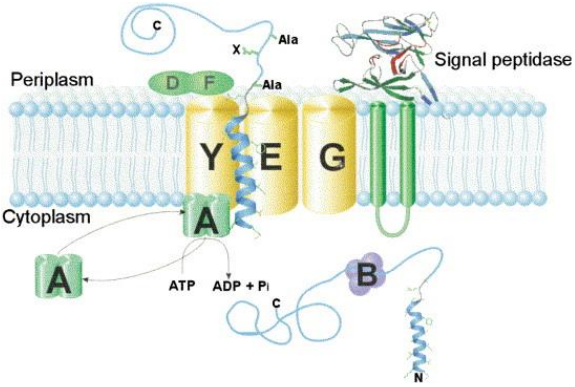


Figure 30. The scheme recapitulates how proteins are translocated across the membrane via the Sec-dependent pathway. Once the preprotein is translocated across the membrane, the type I signal peptidase cleave off the peptide. The figure is

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

7. General discussion and future perspectives

Since their discovery (Nocard and Roux, 1898), *Mycoplasmas* have been interesting species from many different perspectives. Although their small genomes result from genome reduction, they have been considered a model of the simpler organisms found at the origin of complex life. Several diseases can be caused by members of this genus, which moved the interest in these species from something almost philosophical to a practical motivation. For example, animal vaccine companies are interested in generating attenuated *Mycoplasma* strains for vaccination, and also, there is an interest for a vaccine against *M. pneumoniae* in humans. With the advent of what has been called -*omics* technologies, they have become conspicuous target organisms for Systems Biology studies, placing *M. pneumoniae* as one of the most deeply characterized microorganisms. Finally, in the last years, *Mycoplasma* species have become of particular interest for Synthetic Biology. For instance, to make synthetic chromosomes to replace the natural chromosomes removing all genes not essential for life in the laboratory.

This thesis covers, briefly, the multifaceted research interest in mycoplasmas working with a set of different species (*M. agalactiae, M. gallisepticum, M. pneumoniae, and M. feriruminatoris*) and providing advances for the use of species of this genus for practical applications. Chapters 2 and 3 provide tools for some of the historical drawbacks of working with most of these species, such as the low transformation efficiency with transposon vectors and the poor metabolic characterization. [redacted]

Expanding the genome engineering catalog and its applications

Transposon vectors adapted from *S. aureus* were developed in the mid-80s (Lyon et al., 1984), although their efficiency differs substantially between different Mycoplasma species. At the beginning of this present century, a groundbreaking approach was developed in J. Craig Venter Institute with genome transplantation (Lartigue et al., 2007). This allowed bypassing the lack of tools to perform targeted genome editing in *Mycoplasma* by editing their genomes in yeast cells and transferring the modified genomes back to a Mycoplasma cell. Unfortunately, this technique is restricted to a few mycoplasma species as the only mycoplasma cell capable of accepting the transplanted genome is *M. capricolum.* We have tried with *M. pneumoniae* to do chromosome replacement in other species different from Venter's two ones (*M. mycoides* synthetic genome into *M. capricolum* recipient cell), without success. In fact, it has been shown that genome transplantation efficiency is negatively correlated with the phylogenetic distance (Labroussaa et al., 2016). Moreover, it works only in a tiny group of very closely related mycoplasma species. Even after five years of trials, no successful genome transplantation was possible even between two strains of *M. pneumoniae* (FH and M129).

Our group has recently developed an oligo recombineering protocol that makes targeted genome modification of *Mycoplasma* genomes possible. Although oligo recombineering seems to be a portable technology, the reported approach has been described so far only for *M. pneumoniae* (Piñero-Lambea et al., 2020)

This thesis focused on developing efficient transposon vectors for a broad-range of Mycoplasma species. It should be noted that although transposons are a relatively simple tool, they are critical for the initial steps of the research with a novel microorganism, enabling an in-depth characterization in terms of essentiality if the insertion coverage obtained is high enough. They also allowed the introduction of foreign genes. As stated in the introduction, detailed knowledge of essentiality might help develop vaccines, one of the obvious biotechnological applications of Mycoplasmas given the number of diseases caused by them in various farm animals and humans. Specifically, accurate essentiality maps might lead to identifying genes essential for the infection process but dispensable for *in vitro* growth, pinpointing excellent vaccine candidates to test. Indeed, we attempted this identification in *M. agalactiae* with the tools described in Chapters 2 and 3 (data not published). This bacterium uses pyruvate and/or lactate but not glucose because it lacks two critical glycolytic enzymes. The permeases in charge of internalizing pyruvate or lactate in this microorganism are unknown. When using pyruvate as the primary carbon source, the possible pyruvate permease is essential since the administration of a toxic analog of pyruvate (i.e., fluoropyruvate) showed dramatic toxicity, demonstrating it needs to import pyruvate to survive. We introduced growth in the presence of glucose by cloning the two missing glycolytic enzymes, and this engineered strain could survive in the presence of fluoropyruvate (data not published). Using this mutant strain capable of growth with glucose and the *M. agalactiae* WT, we tried to identify the pyruvate transporter by performing essentiality studies in 'conditioned media' (see chapter 3, metabolism) in the presence or absence of glucose and pyruvate. However, not having a defined medium for *M. agalactiae* and using the 'conditional one' resulted in very noisy data that did not allow identification of the pyruvate transporter unequivocally. Having identified this transporter, we would have potentially identified the right vaccination candidate: a KO of pyruvate transporter in a glucose acceptor *M. agalactiae* strain, with the ability to metabolize glucose to grow *in vitro* but not pyruvate, which seems its natural substrate *in vivo*.

The other possibility of obtaining an attenuated strain is to do an essentiality analysis *in vitro* and in the host to identify genes essential for the infection process and not essential for *in vitro* growth. If the mutant library has enough coverage when inoculated in the host and if we can recover enough bacteria after infection, we can compare the find transposon insertions in a gene *in vitro* and *in vivo*. This comparison should identify the most promising genes to inactivate for vaccination: those genes essential for the infection process but dispensable for the *in vitro* growth.

Another critical aspect in developing vaccines of attenuated Mycoplasmas is the need to have a defined medium that could remove the lack of batch reproducibility in fermenters due to heterogeneity of the serum batches and the possibility of viral contamination. First, we need to build a detailed metabolic map with all reactions to have a defined medium. The chemically-defined medium design is, on the one hand, more straightforward in Mycoplasma species due to their genome simplicity but also is complicated because we know that some enzymes have moonlighting activities. Therefore some reactions cannot be assigned to a particular enzyme (i.e., transketolase, Montero-Blay et al., 2020). However, having a map does not mean we can identify which pathways are active and under which circumstances. It is crucial to consider the direction of fluxes and identify the substrate specificity of membrane transporters (there is a significant lack of information for transporters). Chapter 3 of this thesis showed that by combining quantitative proteomics with gene essentiality, one could identify active fluxes and their directions under the conditions tested. The present studies can be performed across different conditions (temperature, thermic or osmotic stress, media composition), and the results obtained can be completely different. It is logical to think that the active metabolic pathways in a cell will vary depending on the medium's composition, the ATP requirements in a particular moment, the medium's pH, etcetera. The analysis done in this chapter serves to identify potential targets for creating attenuated strains and for biotechnological applications like reinforcing or bypassing metabolic fluxes to boost heterologous proteins' expression. The study of the limiting factors when seeking the maximization of protein yield can shed light on which ones are: ATP? Are they cofactors? Can we push metabolism to produce acetate instead of lactate to gain one extra ATP? Can we reflux it to lactate to acquire redox balance? Etcetera.

The tools generated in Chapters 2 and 3 might allow the future development of a defined medium for *M. agalactiae*, opening the door to identifying the pyruvate transporter and the development of a vaccine based on this modification.

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

8. Concluding remarks

We showed for a set of Mycoplasmas how efficient recognition of key transposon elements has a massive impact on transposition efficiency. Then we demonstrated how a newly engineered transposon could be used to test essentiality in four Mycoplasmas encompassing the three Mycoplasmal phylogenetic clades. The new transposon allowed the mapping of the Tn-Seq profile for *M. agalactiae* 7784.

The integration of gene essentiality data, quantitative proteomics data, and homology studies shed light on function and directionality for a set of metabolic pathways, how they intersect, and provide insight into their metabolism's function redundancy two Mycoplasmas: *M. agalactiae* and *M. pneumoniae.*

[redacted]

References

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]