UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE BIOLOGIA VEGETAL

Insights into the transcriptional regulation of *pks1* and *pks15* among *Mycobacterium tuberculosis* complex bacteria

Beatriz Gameiro Ramos

Mestrado em Microbiologia Aplicada

Dissertação orientada por:
Doutora Mónica Vieira Cunha

2018

Acknowledgements

Culminam aqui dois longos anos, de grande aprendizagem pessoal e profissional, que me trazem agora a um dos momentos de maior felicidade e realização que tive, até hoje, na forma desta dissertação. Quero em primeiro lugar, agradecer à minha orientadora Doutora Mónica Vieira Cunha pela oportunidade de abraçar este tema. Durante todo este percurso, proporcionou-me todo o apoio imaterial e financeiro, bem como o estímulo e a liberdade de explorar uma nova abordagem ao tema. Agradeço à Fundação para a Ciência e Tecnologia (Portugal) pelo financiamento deste trabalho através do fundo estratégico ao cE3c (UID/BIA/00329/2013) e do financiamento do projecto PTDC/CVT/117794/2010. Agradeço também ao Doutor Rogério Tenreiro, líder do laboratório de Microbiologia e Biotecnologia, onde desenvolvi este trabalho, não só por me ter recebido no seu laboratório, mas também pela oportunidade de constante aprendizagem que a sua presença representa. Agradeço, igualmente, à Cláudia Ramalho, à Filipa Antunes Silva, à Doutora Lisete Fernandes e à Doutora Ana Tenreiro, por todos os seus conselhos que ajudaram a levar a bom porto o trabalho desenvolvido. Pela sua colaboração e pelos preciosos conselhos para o desenvolvimento deste trabalho, agradeço também à Doutora Rita Zilhão.

Agradeço também aos meus colegas de laboratório, André Pereira, Tiago Baeta, Ana Reis, Inês Santos, Ana Marta Lourenço, Mariana Nascimento, Ana Soares, Catarina Rocha, João Melo e Pedro Teixeira, por todos os momentos de descontração, pela precisão nos horários de almoço e lanche, pela partilha de gomas, bolachas e chocolates (que tanto contribuíram para a minha felicidade) e por tantas outras coisas que contribuíram para que, no fim, fosse possível concluir esta dissertação. Aos meus companheiros de aventura neste mestrado, Ana Vinagre, Alexandra Lança, Ana Fraga, Jorge Figueiredo e Rita Franciosi, um agradecimento por todos os nossos jantares de escárnio e maldizer que nos proporcionaram a oportunidade de falar sobre todos os temas, pessoais e profissionais. Aos meus amigos mais próximos, agradeço pelo contributo que deram a nível pessoal e que me permitiu avançar neste trabalho.

Ao Miguel, agradeço por todo o apoio que me deu ao longo desta dissertação, por todos os nossos almoços e jantares, pelos fins de semana de trabalho, por me ter encorajado sempre a seguir em frente apesar dos contratempos que foram surgindo e por se ter disposto a ser muitas vezes alvo do meu desalento.

À minha família, o meu maior agradecimento por toda a vossa dedicação. Agradeço à minha irmã Inês e aos meus primos Catarina, João Pedro, Henrique e Ricardo por me fazerem viver momentos de genuína felicidade e diversão. Aos meus tios, agradeço pelo incentivo, em particular à minha tia Xana, pelo apoio e preocupação demonstrados desde o primeiro dia desta longa caminhada. Em especial aos meus pais e avós, nunca conseguirei expressar totalmente a gratidão pelo vosso esforço, pelo apoio pessoal, emocional e financeiro que me permitiu seguir o percurso que agora culmina na conclusão desta dissertação.

Insights into the transcriptional regulation of *pks1* and *pks15* among *Mycobacterium tuberculosis* complex bacteria

Beatriz Gameiro Ramos

2018

This thesis was fully performed at Microbiology and Biotechnology Lab (LMBBioISI|Bugworkers|TecLabs) and Centre for Ecology, Evolution and Environmental Changes (cE3c), under the direct supervision of Professor Mónica Vieira Cunha, in the scope of the Master in Applied Microbiology of Faculdade de Ciências da Universidade de Lisboa.

Abstract

The *Mycobacterium* genus belongs to the Actinobacteria phylum, is the single member of the Mycobacteriaceae family and includes more than one hundred and seventy species. Mycobacteria are defined as acid-fast actinomycetes that usually form slightly curved or straight non-motile rods. Its acid-fastness is due to the presence of mycolic acids, an unique class of lipids that can only be found in the cell wall of species from *Mycobacterium* genus. This genus is also characterized by a high G+C content (61-71%). The species from *Mycobacterium* genus can be divided into two groups, the first includes slow-growing species like the pathogens *Mycobacterium tuberculosis* (*Mtb*), *Mycobacterium bovis* (*Mb*) and *Mycobacterium leprae*, and the second that includes fast-growing species, mostly opportunists or non-pathogens, like *Mycobacterium smegmatis*. *Mycobacterium tuberculosis* complex (MTC) is constituted by *Mtb* alongside with other genetically related species of mycobacteria, such as *Mycobacterium canettii, Mycobacterium africanum, Mycobacterium microti, Mb, Mycobacterium caprae* or *Mycobacterium pinnipedii*, each one best adapted to particular host species. Amongst the members of the MTC, *Mtb* is the emblematic etiological agent of human tuberculosis (TB). In 2015, TB was amongst the top 10 causes of death worldwide, even above HIV/AIDS as a leading cause of death by infection. Even though TB mortality and incidence rates are decreasing in most parts of the world, the disease remains spread worldwide, with particular incidence in Africa, Asia, and Eastern Europe. Although rare or sporadic in developed countries, infection by *Mb* in humans typically occurs by inhalation of aerosols in individuals in close contact with infected animals or by the ingestion of contaminated unpasteurized milk. Infection by *Mb* in humans is clinically indistinguishable from infection by *Mtb*.

The clinical evolution of mycobacterial infection depends on the virulence of the infecting bacterium and on the ability of the particular host to limit bacterial replication. In an immunocompetent host, immune system cells constraint the infection in a structure called granuloma. While inside granulomas, mycobacteria experience a set of stress conditions, like hypoxia, micronutrient starvation and acidic pH, and develop a dormant, non-replicating state that can reactivate upon host immunosupression.

One of the most relevant characteristics of mycobacteria is its cell wall, composed by three major structures. To the cell wall core are esterified mycolic acids, long-chain fatty acids specific of mycobacterial cell wall that are directly related to cell viability and impermeability. Amid the *Mycobacterium*-specific components of several pathogenic mycobacteria, namely *Mtb* and *Mb*, are two related families of glycosylated lipids: diphthioceranates (DIP) and phthiocerol dimycocerosate (PDIM) and its variant phenolic glycolipids (PGL). PGL are known to be associated with impermeability of the cell wall, phagocytosis, defense mechanisms against nitrosative and oxidative stress and, theoretically, to the ability of mycobacteria to form biofilms. The genes related to the biosynthesis of PGL belong to the class of polyketide synthases (PKS).

In this work, we focused on the role of *pks1* and *pks15*, two genes known to be involved in the biosynthesis of PGL. Previous studies from our research group have already described the nucleotide and aminoacid diversity of those genes and encoded proteins, along with their microevolutionary history, and have also assessed the association between biofilm formation and the profile of membrane lipids, in different stress conditions. Transcriptional analysis of *pks1* and *pks15* genes was also performed by RT-qPCR and results showed increased expression under nitrosative and oxidative stress, including in the presence of ascorbic acid. Since there is bare regulatory information for those two genes, this thesis aims were to: [1] define whether or not *pks1* and *pks15* are co-transcribed as polycistronic transcription units using an experimental approach targeting the construction of transcriptional fusion vectors and promoter strength

inference; and [2] explain the function of *pks1* through the construction of a transposon-free knock-out mutant relying on phage mediated mutagenesis and mycobacterial recombineering experimental methodologies; and [3] identify the regulatory pattern responsible for controlling *pks1* and *pks15* transcription using a genome-wide approach based on transcriptome data (RNA-seq) available at public databases, like *Tuberculist*, *National Center for Biotechnology Information (NCBI),TB Database* and *MTB Network Portal*.

Cloning of putative regulatory regions was individually performed in pJET1.2/blunt for *pks1*, *pks15* and *fadD22*, the minimal structure proposed for this operon, and was successful for all three targets. However, subsequent subcloning in the mycobacterial shuttle vector pSM128 was ineffective, despite several attempts. Another experimental strategy that did not work favourably was the attempt to generate a knock-out mutant of *pks1*, neither by phage-mediated mutagenesis or mycobacterial recombineering. We successfully made a recombineering strain with pJV53, although the attempts to electroporate the allelic exchange substrate (AES) containing the upstream and downstream regions of *pks1* for homologous recombination were unsuccessful. By phage-mediated mutagenesis, we were able to generate several transductants but, due to the high dimension of the desired DNA construct (over 50 kb), we were not able to successfully confirm the ligation of the AES cloned in a cosmid with phasmidic DNA.

Data collected from different databases point out that *pks1* and *pks15* may be transcriptional units of an operon composed by three to six genes (*lppX*, *pks1*, *pks15*, *fadD22*, *Rv2949c* and *fadD29*), all involved in the biosynthesis of the phenolptiocerol moiety of PGL with the exception of *lppX*. It has also been proposed, by *in silico* analysis, that both *pks1* and *pks15* are positively regulated by *sigK* and negatively by *sigE*.

By clustering the transcriptome data from RNA-seq for *Mtb*, in a robust set of 25 growth conditions corresponding to 60 experiments from five stress datasets, in this work it was possible to relate *pks1* with *fadD22*, *Rv2949c*, *fadD29*, *pks6*, *pks12* and *pks9*, gathering evidence supporting that the first three genes are part of a polycistronic structure. Clustering analysis particularly grouped *pks1*, *fadD22*, *Rv2949c* and *fadD29* with correlation values above 0.9, also including some of *pks1* paralogs, namely *pks6*, *pks12* and *pks9*. The *pks1*, *Rv2949c*, *fadD22* and *fadD29* genes were also highly correlated with *sigC*, *sigG* and *sigK* sigma factors (correlation values above 0.8). Despite RNA-seq information is scarcer for *Mb*, we analysed two growth conditions corresponding to 17 experiments from two stress datasets, finding a significant cluster (correlation value above 0.85) composed by *lppX*, *pks15/1*, *fadD22*, *Rv2949c*, *fadD29* and also *Mb0429c* and *pks13*.

By analysing the differential expression of *lppX*, *pks1*, *pks15*, *fadD22*, *Rv2949c* and *fadD29* using large RNA-seq datasets, it was then possible to define in which conditions are those genes positively or negatively regulated in *Mtb*. We confirm that the selected genes of interest are up-regulated in acidic pH and down-regulated at stationary phase, under hypoxia and dormancy, and at both low and high iron concentrations. Down-regulation of these genes in these latter conditions is consistent with a non-replicating state and reduction of the synthesis of cell wall components by mycobacteria inside the host, as the bacterium essential cellular processes become progressively affected by host immune response.

Using differential expression analysis, we were also able to confirm that *sigK* shares the expression profile with the selected genes of interest in 80% of the analysed conditions with significant fold-changes; the same percentage was also verified for *sigD*. On the contrary, in approximately 78% of the conditions under analysis, *sigE* expression profile is dissimilar with those of the selected panel of genes with significant fold-changes, as well as *sigB*.

Finally, gathering the previously published information and the regulatory data that we were able to collect and analyse in this work, we propose a regulatory model for *pks1* and *pks15*. In

this predictive model, we define a conservative approach in which only genes sharing coherent expression profiles and functional similarity of the polycistronic structure are integrated, specifically *pks1*, *pks15* and *fadD22*. Based on differential expression analysis, we select a set of four σ factors (σ^D and σ^K, and σ^B and σ^E) for which we find experimental support to the regulation of the expression of *pks1*, *pks15* and *fadD22*. The genes encoding σ^D and σ^K factors were shown to be down-regulated under hypoxia and dormancy, as well as in stationary phase. On the contrary, the genes encoding σ^B and σ^E factors are up-regulated in the same conditions, being that *sigB* was previously shown to be up-regulated under hypoxia. Both σ^K and σ^E were previously predicted to regulate the genes belonging to the polycistronic structure under hypothesis and our findings in the context of the present work give support to that prediction. However, we propose that σ^D and σ^B are also involved in the regulation of *pks1*, *pks15* and *fadD22*. While σ^D and σ^K appear to positively regulate the selected genes of interest and putatively belong to the lower level of σ factors regulation, we find evidence that σ^B and σ^E factors negatively regulate *pks1*, *pks15* and *fadD22* at an upper level.

In the future, completing the experimental component originally foreseen in this work, such as promoter activity testing under different growth conditions and construction of a loss-of-function *pks1* mutant, will significantly contribute to refine knowledge on the regulation of *pks1* and *pks15* transcription, clarify how their transcription is articulated with global gene expression networks and thus further enlighten the biological role of these genes products in the mycobacterial cell.

Resumo

O género *Mycobacterium* pertence ao filo Actinobacteria, sendo o único membro da família Mycobacteriaceae e tendo mais de 170 espécies descritas. As micobactérias têm como principais características o seu elevado conteúdo em G+C (61-71%) e a propriedade de álcool-ácido resistência, atribuída à presença de ácidos micólicos na sua parede celular. As bactérias deste género podem ser agrupadas de acordo com o seu crescimento, estando no primeiro grupo incluídas as espécies de crescimento-lento, tais como *Mycobacterium tuberculosis* (*Mtb*), *Mycobacterium bovis* (*Mb*) e *Mycobacterium leprae* e, num segundo, as espécies de crescimento-rápido, maioritariamente oportunistas e não-patogénicas, como *Mycobacterium smegmatis*. Além desta divisão das espécies pelo seu crescimento, estas são também agrupadas segundo a sua semelhança genética, nomeadamente no complexo *Mycobacterium tuberculosis* (MTC), composto por *Mtb* e por outros ecótipos que partilham, ao nível das sequências nucleotídicas do genoma, mais de 99% de semelhança. Os diferentes ecótipos são identificados de acordo com marcadores moleculares determinados, nomeadamente 14 regiões de diferença (RD's) e seis regiões de eliminação (RvD's). Entre os 20 marcadores moleculares, destacam-se RD9 e TbD1. A linhagem que apresenta a eliminação da RD9 inclui *M. africanum*, juntamente com outras espécies presentes em reservatórios animais, tais como, *M. microti*, *M. pinnipedii*, *M. bovis*, *M. caprae*, *M. mungi* e *M. orygis*. A eliminação de TbD1 carateriza um grupo de linhagens designadas de estirpes "modernas" de *Mtb*, enquanto os restantes membros de MTC não apresentam esta deleção.

A espécie *Mtb* é o agente etiológico paradigmático da tuberculose (TB) humana. Ao nível mundial, a TB encontra-se no top 10 de causas de morte por infeção, acima da infeção por HIV/SIDA. Apesar de, nos últimos anos, se ter verificado uma diminuição das taxas de mortalidade e incidência, a doença mantém-se distribuída mundialmente, com principal incidência em Africa, na Ásia e na Europa oriental. A doença tem expressão, maioritariamente, pulmonar, mas 15% dos casos notificados em 2015 à Organização Mundial de Saúde (OMS) representam TB extrapulmonar. Em 2015, foram também notificados 480 mil casos de TB multirresistente (MDR-TB). A MDR-TB é prevalente em doentes previamente expostos a tratamentos anti-TB de forma irregular, favorecendo a seleção de mutantes resistentes. *Mb*, sendo um ecótipo adaptado a ungulados e carnívoros, é um agente zoonótico de TB em humanos, sendo que a infeção não é distinguível clinicamente da infeção causada por *Mtb*. Tipicamente, esta infeção ocorre por inalação de aerossóis em pessoas em contacto próximo com animais infetados ou por ingestão de leite não pasteurizado.

A extensão da infeção por *Mtb* é dependente de dois parâmetros, a virulência da estirpe infetante e a competência da resposta imune do hospedeiro. As micobactérias patogénicas apresentam-se largamente adaptadas à sobrevivência dentro de macrófagos. Graças à inibição da fusão dos endossomas com os lisossomas, as micobactérias evadem-se dos mecanismos bactericidas que ocorrem nos lisossomas. A infiltração linfocitária e a ação concertada dos componentes do sistema imunitário no local da infeção limitam a disseminação das micobactérias numa estrutura designada granuloma. Durante a permanência dos bacilos no granuloma, estes são sujeitos a diversos fatores de stress, tais como hipoxia, escassez de nutrientes e pH acídico, que conduzem ao desenvolvendo de um estado de latência, não replicativo. O metabolismo micobacteriano é, no entanto, reativado mediante a supressão do sistema imunitário do hospedeiro.

A parede celular das micobactérias é uma das suas principais características, sendo constituída por três estruturas, incluindo polissacáridos capsulares. O centro da parede celular é composto por peptidoglicano em ligação covalente com arabinogalactano que, por sua vez, está esterificado aos ácidos micólicos. Na parede estão também presentes glicolípidos, lipoglicanos e lipoproteínas. Entre os componentes específicos de *Mycobacterium*, estão os dimicocerosatos

de ftiocerol (DIM) e os glicolípidos fenólicos (PGL), os quais têm sido associados à impermeabilidade da parede celular, à fagocitose, aos mecanismos de defesa contra os stresses oxidativo e nitrosativo e, teoricamente, à capacidade das micobactérias formarem biofilmes. A maioria dos genes responsáveis pela biossíntese de DIM e PGL estão localizados no *locus* DIM + PGL, nomeadamente os genes *fadD26*, *ppsA-E*, *fadD28*, *lppX*, *pks15*, *pks1*, *fadD22*, *Rv2949c* e *fadD29*. Entre estes, os genes *pks1* e *pks15* estão documentados como estando envolvidos na síntese de PGL, sendo que constituem uma única grelha de leitura aberta (*pks15/1*) em estirpes produtoras de PGL. Apesar de se conhecer a associação de *pks15/1* a esta via biossintética, ainda não se encontra esclarecida a sua regulação transcricional.

Tal como noutros procariotas, a transcrição em micobactérias é regulada pela associação de fatores σ à RNA polimerase, que assim formam uma holoenzima funcional na ativação da expressão génica. Cada fator tem afinidade para sequências promotoras específicas, regulando a expressão de genes alvo. Os 13 fatores σ descritos em micobactérias têm sido associados a diferentes condições de crescimento, nomeadamente o factor σ^A que assegura a expressão constitutiva de muitos genes do genoma micobacteriano, e o factor σ^B, associado à resposta geral ao stress. A função destes fatores tem sido estudado nos últimos anos por recurso à análise dos respetivos níveis de expressão e, mais recentemente, através de mutantes de eliminação, uma vez que o seu nível de expressão pode não representar diretamente a condição em que o fator efetivamente se associa à RNA polimerase.

Neste trabalho, procurou-se esclarecer aspetos regulatórios e funcionais dos genes *pks1* e *pks15*, que se sabe estarem envolvidos na biossíntese de PGL. Estudos prévios do nosso grupo de investigação analisaram a diversidade nucleotídica e aminoacídica destes genes, bem como a sua historia microevolutiva. Além disso, analisou-se também previamente a associação entre a formação de biofilme e o perfil dos lípidos de membrana de micobactérias patogénicas, em diferentes condições de stress nitrosativo e oxidativo, incluindo na presença de ácido ascórbico. O aumento de expressão destes genes nessas condições foi demonstrado por RT-qPCR. No entanto, uma vez que a informação disponível sobre os aspetos mecanísticos que regulam a expressão destes genes é reduzida, os objetivos do presente trabalho foram especificamente: [1] aferir se os genes *pks1* e *pks15* se encontram numa estrutura policistronica, através da construção de vetores de fusão transcricional e inferência da localização exata e condições de ativação do promotor; [2] estudar a função do gene *pks1* através da construção de mutante de eliminação de *pks1* por mutagénese mediada por fagos e por *recombineering;* e [3] inferir o padrão de regulação transcricional dos genes *pks1* e *pks15* pela análise de dados em larga escala do transcritoma (RNA-seq).

A clonagem das regiões promotoras dos genes *pks15*, *pks1*, *fadD22* foi inicialmente realizada no vetor pJET1.2/blunt, sendo bem-sucedida para os três alvos. Pelo contrário, a subclonagem dos mesmos alvos no vetor de fusão transcricional pSM128, replicativo em *Escherichia coli* e micobactérias, não foi concluída com sucesso, apesar das várias tentativas de otimização do rácio inserto:vector e das condições de ligação. Da mesma forma, a construção do mutante de eliminação do gene *pks1* pelas duas abordagens independentes previstas não foi concluída. Pela metodologia de *recombineering*, foi possível gerar uma estirpe recombinante de *Msm* contendo o plasmídeo pJV53, sendo que não foi possível recuperar nenhuma estirpe contendo o substrato de troca alélica. Através da mutagénese mediada por fagos, o constrangimento encontrou-se na dimensão do DNA plasmídico que se pretendia construir para transdução, uma vez que a clonagem do substrato de troca alélica construído em cosmídeo no fagemídeo derivado do fago lambda gera uma construção que ultrapassa os 50 kb, ultrapassando também o limite máximo de extração da maioria dos sistemas comerciais e de métodos *in-house*, dificultando enormemente a recuperação do DNA plasmídico e a confirmação da obtenção do substrato transdutor.

Para inferir a regulação transcricional dos genes em estudo, recolheu-se a informação disponível em diversas bases de dados, tais como *Tuberculist, National Center for Biotechnology Information (NCBI), TB Database* e *MTB Network Portal*. Além destes dados, a pesquisa do local de ligação do ribossoma (RBS) foi realizada através do *Prokaryotic Dynamic Programming Genefinding Algorithm* (PRODIGAL). A análise de sintenia foi também realizada para os genes *pks1, pks15* e *fadD22* através do *SyntTax* (*Prokaryotic Synteny & Taxonomy Explorer*). Para tentar definir o padrão e a mecanística da regulação transcricional, foi realizada uma análise de dados em larga escala do transcritoma, obtidos por RNA-seq, constituindo um conjunto de 27 condições experimentais (25 para *Mtb* e 2 para *Mb*) e um total de 77 replicados (60 para *Mtb* e 17 para *Mb*), para um grupo de 52 (*Mtb*) e 50 (*Mb*) genes codificando sintetases de poliquétidos, fatores σ e membros dos módulos de *bicluster* 0211 e 0490 definidos pelo algoritmo *cMonkey* (disponíveis no *MTB Network Portal*).

A informação recolhida de bases de dados sugere que os genes *pks1* e *pks15* poderão formar uma unidade transcricional composta por 3 a 6 genes, localizados tanto a montante do gene *pks15* como a jusante do gene *pks1*. Destes 6 genes (*lppX*, *pks1*, *pks15*, *fadD22*, *Rv2949c* e *fadD29*), apenas o gene *lppX* não está envolvido na síntese de fenoltiocerol. Dados anteriores baseados na análise *in silico* indicam que os genes *pks1* e *pks15* são putativamente regulados positivamente por *sigK* e negativamente por *sigE*. No presente trabalho, a análise dos dados de transcritoma indicou que, para *Mtb*, é possível associar num único *cluster*, com pelo menos 85% de semelhança, aferida pelo método de aglomeração UPGMA, os genes *pks1, fadD22, Rv2949c, fadD29, pks6, pks12* e *pks9*. Para *Mb*, devido à existência de um grupo menor de dados, um *cluster* com o mesmo nível de semelhança inclui os genes *lppX, pks15/1, fadD22, Rv2949c, fadD29* e, ainda, os genes *Mb0429c* e *pks13*. Os dados de transcritoma sugerem ainda uma associação entre os membros da estrutura policistrónica putativa e os genes que codificam para os fatores σ^C, σ^G e σ^K, aferida por um coeficiente de correlação de Pearson superior a 0,8.

Com os dados de RNA-seq, foi também possível estabelecer uma analise de expressão diferencial que permite identificar, por comparação, em que condições os membros da estrutura policistronica putativa estão sobre e subexpressos. Dentro do lote de condições analisadas, que inclui pH acídico, fontes de carbono alternativas, diferentes fases do crescimento bacteriano, exposição a diferentes concentrações de ferro, hipoxia e latência, apenas foi verificada a sobre-expressão dos genes alvo na amostra crescida em pH acídico, quando comparada com a amostra crescida em pH neutro. A adição de piruvato como fonte de carbono alternativa ao meio de cultura não promoveu alterações significativas na expressão dos genes alvo, quando comparando com as amostras crescidas unicamente em glicerol. A comparação das amostras recolhidas em fase estacionária do crescimento com as amostras recolhidas em fase exponencial, no mesmo meio de cultura, revelaram a subexpressão dos genes alvo em fase estacionária, tal como previsto. Tanto para as amostras crescidas em alta como em baixa concentração de ferro, os genes alvo aparentam ser subexpressos por comparação com a amostra crescida em condições de crescimento padrão *in vitro*. Tanto para as amostras recolhidas da cultura em hipoxia, como para as amostras recolhidas da cultura em latência, os genes alvos apresentam elevada subexpressão. Para *Mb*, analisaram-se ainda culturas crescidas em condições de carência nutricional, revelando o mesmo padrão apresentado em *Mtb* para culturas em hipoxia e latência. A combinação da informação resultante da análise de *clustering* e da análise de expressão diferencial parece apontar para que o gene *lppX*, de função parcialmente desconhecida, tenha uma regulação transcricional mais divergente da dos restantes genes alvo. Confirma-se, assim, neste trabalho que os genes em estudo são regulados positivamente em pH ácido e subexpressos em fase estacionária, sob hipoxia e latência e em concentrações baixas e altas de ferro. A regulação negativa destes genes nestas últimas condições é consistente com o estado não replicativo das micobactérias no seio do hospedeiro e a subsequente redução da síntese de

componentes da parede celular à medida que os processos celulares essenciais da bactéria são afetados pela resposta do sistema imunitário do hospedeiro.

A análise de expressão diferencial baseada no transcritoma permitiu também confirmar que o gene *sigK* partilha o padrão de expressão com os genes de interesse em 80% das situações analisadas que apresentam diferenças de expressão significativas; a mesma percentagem foi também verificada para o gene *sigD*. Pelo contrário, em aproximadamente 78% das condições analisadas, com diferenças de expressão significativas, o gene *sigE* apresenta um perfil diferente do encontrado nos genes de interesse, tal como acontece com *sigB*.

Compilando a informação recolhida das bases de dados e a informação regulatória obtida através da análise de transcritoma, propõe-se neste trabalho um modelo de regulação para os genes *pks1* e *pks15*, tendo por base uma abordagem conservativa em que se selecionou apenas os genes com total semelhança funcional e perfil transcricional. Foram, assim, selecionados os genes *pks1*, *pks15* e *fadD22*, bem como os fatores σ^D, σ^K σ^B e σ^E. Os genes que codificam os fatores σ^D e σ^K encontram-se regulados negativamente em condições de hipoxia e latência, bem como na fase estacionária, sendo que o padrão contrário é verificado para os genes que codificam os fatores σ^B e σ^E. Além dos fatores σ^K e σ^E, previamente propostos como reguladores dos genes de interesse, propõe-se também que os fatores σ^B e σ^D estejam envolvidos na regulação de *pks1*, *pks15* e *fadD22*.

No futuro, será da maior relevância concluir as metodologias moleculares iniciadas no presente trabalho, tais como o estudo da atividade do promotor sob diversas condições de stress que mimetizam o ambiente do hospedeiro e a construção de um mutante de eliminação do gene *pks1*. Estas abordagens poderão contribuir para refinar o conhecimento da regulação transcricional dos genes *pks1* e *pks15*, clarificar aspetos mecanísticos e a forma como a transcrição destes genes é articulada com redes de expressão globais e, assim, contribuir para o esclarecimento da função biológica exercida pelo produto destes genes na parede celular das micobactérias.

Table of Contents

List of Figures and Tables

Abbreviations

µF – Microfarads

µg – Microgram

µl – Microliter

ºC – degrees Celsius

7H10 – Middlebrook 7H10 agar

7H9 – Middlebrook 7H9 broth

ACP – Acyl Carrier Protein

AES – Allelic Exchange Substrate

AT – Acyltransferase

BAM – Binary Sequence Alignment Map

bp – base pair

cm – Centimeter

cPCR – colony PCR

DH – Dehydratase

DIP – Diphthioceranates

DMSO - Dimethyl sulfoxide

E. coli – *Escherichia coli*

ER – Enoyl Reductase

FPKM – Reads Per Kilobase of exon model per Million mapped reads

g – gram

h – Hour

HIV/AIDS – Human immunodeficiency virus/Acquired immune deficiency syndrome

IFN-γ – Interferon gamma

IL-12 – Interleukin 12

kb – Kilo base pair

KR – Keto Reductase

KS – Ketoacyl synthase

kV – Kilovolts

l – Liter

LA – Luria-Bertani agar

LB – Luria-Bertani broth

ManLam – Lipoarabidomannan

Mb -Mycobacterium bovis

MDR-TB – Multidrug-resistant tuberculosis

min – Minutes

ml – Mililiter

mM – Milimolar

Msm – Mycobacterium smegmatis

Mtb – Mycobacterium tuberculosis

MTC – *Mycobacterium tuberculosis* complex

NCBI – National Center of Biotechnology Information

NEB – New England Biolabs

ng – Nanogram

NK – Natural Killer cells

o/n – overnight

OD_{600} – Optical Density at 600 nm

PAMP – Pathogen Associated Molecular Patterns

PCR – Polymerase Chain Reaction

PDIM – Phthiocerol Dimycocerosates

pDNA – plasmid DNA

PGL – Phenolic Glycolipids

p-HBA – p-hydroxybenzoic acid

PIM – Mannosylated Phosphotidylinositol

PRODIGAL – Prokaryotic Dynamic Programming Genefinding Algorithm

PRR – Pattern Recognition Receptors

RBS – Ribosomal Binding Site

RDs – Regions of Difference

Rpm – Rotations per minute

rSAP – Shrimp Alkaline Phosphatase

RT – room temperature

RvDs – Regions of Elimination

s – Seconds

SNPs – Single Nucleotide Polymorphisms

SOC - Super Optimal broth with Catabolite repression

TB – Tuberculosis

TH1 – T-helper 1 cells

TLRs – Toll-like Receptors

TNF- α – Tumour Necrosis Factor α

TSS – Transcription Start Site

UPGMA - Unweighted Pair Group Method with Arithmetic Mean

V – Volts

α – alpha

β – beta

γ – gamma

σ – sigma

Ω – Ohm

ω – omega

1. Introduction

1.1 Introducing the *Mycobacterium* genus

Mycobacteria are defined as aerobic, acid-fast actinomycetes that usually form slightly curved or straight non-motile rods. Its acid-fastness is due to the presence of mycolic acids, an unique class of lipids that can only be found in the cell wall of species from *Mycobacterium* genus[1, 2]. Those species are also characterized by an high G+C content (61-71%) and an optimal growth temperature ranging from 25 to 45°C[3]. The *Mycobacterium* genus belongs to the Actinobacteria phylum, is the single member of the *Mycobacteriaceae* family and has more than one hundred and seventy species described. Those can be divided into two groups, the first includes slow-growing species, like the pathogens *Mycobacterium tuberculosis* (*Mtb*), *Mycobacterium bovis* (*Mb*) and *Mycobacterium leprae*, and the second that includes fast-growing species, mostly opportunists or non-pathogens, like *Mycobacterium smegmatis*. *Mycobacterium tuberculosis* complex (MTC) is constituted by *Mtb* alongside with other genetically related species of mycobacteria, like *Mycobacterium canettii*, *Mycobacterium africanum*, *Mycobacterium microti*, *Mb*, *Mycobacterium caprae*, *Mycobacterium pinnipedii*, *Mycobacterium mungi* and *Mycobacterium orygis*, each one adapted to preferential hosts[1, 4]. *Mtb*, *M. africanum* and *M. canetii* are usually hosted by humans, *M. microti* by rodents, *M. caprae* by goats, *M. pinnipedii* by marine mammals, *M. mungi* and *M. orygis* by mongooses and antelopes, respectively, and *Mb* by cattle and other species[5, 6].

Typically, at the genome level, MTC members present a nucleotide sequence similarity above 99% and exhibit a clonal structure, with sparse evidence of exchange of chromosomal DNA, except for a few members and *M. canettii* that present smooth colony morphology and for which there is evidence of recombined DNA segments present in their chromosomes[7]. Different ecotypes are identified according to fixed molecular markers, including single nucleotide polymorphisms (SNPs) and deletions. There are 20 variable regions for differentiation of ecotypes, of which 14 are regions of difference (RDs) and six are regions of elimination (RvDs)[8]. Amongst those variable regions are two major phylogenetic markers that enable clustering of the strains, namely RD9 and TbD1. The lineage deleted for RD9 includes *M. africanum*, a human pathogen mostly found in West Africa, and other species that are maintained in animal reservoirs, such as *M. microti*, *M. pinnipedii*, *M. bovis*, *M. caprae*, *M. mungi* and *M. orygis*. *Mtb* TbD1 deleted region is also a significant marker that is absent in a cluster of lineages named as "modern" *Mtb* strains, while it is present in the remaining MTC members[9]. This deletion implies the absence of a member of the *mmpL* gene family that encodes mycobacterial membrane proteins that are dedicated to the transport of several cell wall components[9]. Amongst the TbD1-deleted strains are Lineage 4 (also known as Euro-American), known to have a broad distribution in Europe, America and also in Africa and the Middle-East; Lineage 2 (including Beijing family), mostly spread in East Asian countries and Lineage 3, with a narrow distribution, occurring in East Africa and in Central and South Asia. Besides the animal-adapted strains, other strains that do not present TbD1 deletion are comprised in Lineage 1 (also known as Indo-Oceanic), occurring around the Indian Ocean and the Philippines, Lineage 5 (also known as *M. africanum* West African 1), and Lineage 6 (*M. africanum* West African 2), both restricted to Western African countries and Lineage 7 (Ethiopian lineage) (Fig. 1.1).

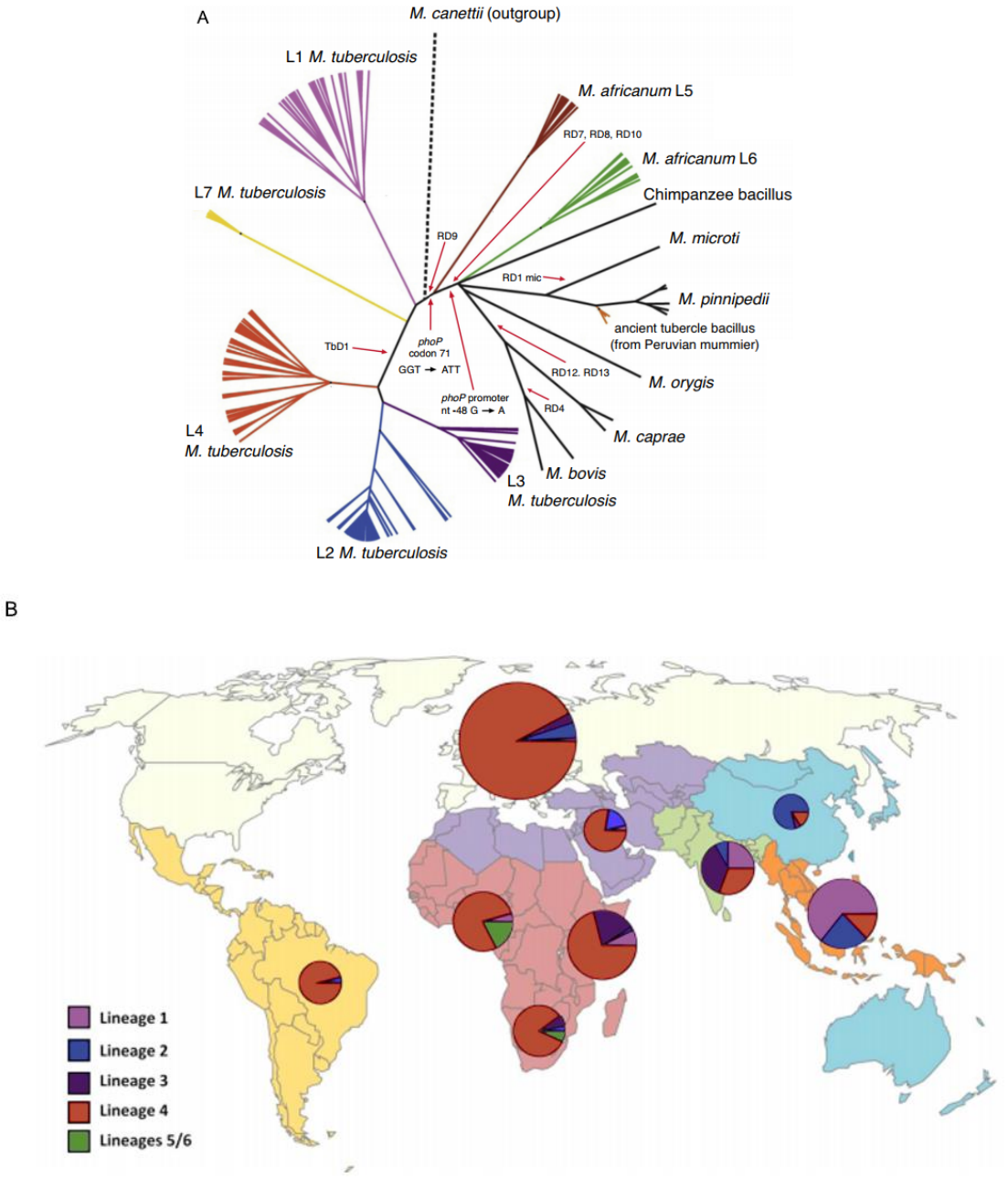

Figure 1.1 – Phylogenetic relationships of MTC lineages and their global distribution. A – Representation of the phylogenetic relationships of MTC lineages and information on deletion points of selected RDs. Adapted from Orgeur and Brosch (2018). B – Distribution of the main phylogenetic lineages worldwide, according to the geographic origin of the patients. Adapted from Fenner *et al.* (2013).

1.2 Epidemiology of tuberculosis worldwide

Current knowledge postulates that *Mtb* colonizes approximately one third of the human population worldwide, being also capable to infect other mammal species that are in direct contact with humans. In humans, tuberculosis (TB) is an infectious disease caused by the bacillus *Mtb*, that represents a major public health concern. The disease is spread by droplets containing bacilli expelled by coughing[10].

In 2015, *Mtb* was estimated to cause 10.4 million new cases of TB and 1.4 million deaths, being amongst the top 10 causes of death worldwide, even above HIV/AIDS as a leading cause of death by infection. Even though TB mortality and incidence rates are decreasing in most parts of the world, the disease remains distributed worldwide, with particular incidence in Africa, Asia, and Eastern Europe. Portugal is still the country with the highest TB incidence in Western Europe[10, 11]. Although TB is known to be mainly a pulmonary disease, it can also affect other body sites. In agreement, 15% of the 6.1 million incident cases notified to World Health Organization (WHO) in 2015 represented extrapulmonary TB[10]. There were also an estimated 480 thousand new cases of multidrug-resistant TB (MDR-TB). MDR-TB is prevalent in patients that were previously exposed to anti-TB treatments in an interrupted, disadjusted or irregular manner, favouring the selection of drug-resistant mutants. Most immunocompetent infected people contain the bacilli inside granulomas after a balance between the bacilli and the immune system is achieved. CD4+ T cells greatly affect this process as the production of cytokines induces activation of macrophages. In HIV-patients, the CD4+ T cells count declines and the ability of the host to maintain the bacilli circumscribed to a few infected macrophages declines as well[12]. In 2015, from the TB patients with a reported HIV status, 6% were coinfected[10].

Infection by *Mb* in humans happens sporadically, is more frequent in under-developed countries, and it typically occurs by inhalation of aerosols or the ingestion of unpasteurized milk, being clinically indistinguishable from infection by *Mtb*. Even though there is no routine reporting of zoonotic TB cases in most countries where it is endemic, it is estimated that, in 2015, there were approximately 149 thousand cases[10]. Besides the zoonotic risk, bovine tuberculosis is also highly relevant in the economy of livestock farming, once it directly influences animal productivity, trade of animal products and infected animals have to be slaughtered, with subsequent economic losses[13, 14].

1.3 Mycobacteria-host interaction

Mycobacterial infection can cause different consequences depending on two main parameters, the virulence of the infecting bacteria and the resistance of the particular host[15]. Pathogenic mycobacteria are widely adapted to survive intracellularly within the macrophages. By preventing the fusion of the phagosomes with lysosomes, the bacterium is able to avoid the bactericidal mechanisms that operate within lysosomes and hence processing and presentation of its pathogen associated molecular patterns (PAMP) to the host's immune system. Cells of the host's immune system are able to recognize those PAMP through their pattern recognition receptors (PRR)[16-18]. There are plenty PRR already known to play a role in human macrophage phagocytosis of *Mycobacterium* spp., such as collectins, C-type lectins and toll-like receptors (TLRs). On the other side, PAMP for pathogenic mycobacteria include CpG DNA, Lipoarabinomannan (ManLam) and mannosylated phosphotidylinositol (PIM)[18].

Upon the entrance of *M. tuberculosis*, dendritic cells become activated by TLRs signals, initiating the activation of T cells. *Mtb* ligands for TLRs will promote inflammation, characterized by releasing of chemokines and pro-inflammatory cytokines, like interleukin 12 (IL-12) and interleukin 18 (IL-18), which induce natural killer (NK) cells and biase the immune response towards T helper 1 cells (Th1) that will lead to the secretion of interferon gamma (IFN-γ). Macrophages will be activated by the secreted IFN-γ and become more competent effector cells, being able to express microbiocidal substances and cytokines, amongst which is tumour necrosis factor α (TNF-α) that plays a central role in mycobacterial infection control and granuloma formation[17, 19].

Granuloma consists of a central necrotic core, containing liquefied tissue, surrounded by layers of macrophages, dendritic cells, lymphocytes, neutrophils and fibroblasts. This structure is beneficial for the host, since it preserves the infection restrained to localized regions, preventing bacterial spreading. Although mycobacteria are exposed in the granuloma to several stresses induced by the host, like starvation, reactive oxygen and nitrogen intermediates, or hypoxia, they are able to remain on a metabolic latent phase, less susceptible to immune system aggressions[16, 20]. A mature granuloma typically represents an area of infection characterized by intense immunological activity wherein activated macrophages present mycobacterial PAMPs to T-lymphocytes, inducing them to produce a variety of chemokines and cytokines[16]. When compromised, for example due to HIV infection, the immune system is unable to contain the bacilli in the granulomas and active disease is then potentially triggered[16].

1.3.1 Mycobacterial growth under host induced stress

Mycobacteria experience a set of stress conditions inside the granuloma, like hypoxia, nitrosative stress, micronutrient starvation and acidic pH.

Oxygen deprivation, along with the presence of NO and CO generated by immune response mediators, inhibit aerobic respiration and induce the expression of the DosR regulon[21]. This regulon is composed by three components, the two histidine kinases DosS and DosT and the response regulator DosR[21]. Either DosS or DosT autophosphorilate one of their histidines and thus phosphorylate an aspartate residue of DosR, promoting the binding of DosR to DNA in the upstream region of responsive genes to hypoxia[21, 22], thus triggering dormancy.

Iron is an essential micronutrient in most aerobic bacteria and plays a relevant role on several biological processes, such as electron transport and redox reactions. It can be harmful at high concentrations once it mediates the formation of hydroxyl radicals that can damage both DNA and proteins[23, 24]. Pathogenic bacteria face another challenge, since mammalian hosts limit the available amount of iron. Inside macrophages, iron concentrations are low, varying between 1 and 10 ng/ml. Those low concentrations are maintained, even though there is a high flux of this metal inside macrophages due to erythrocytes' destruction[23, 24]. To maintain the iron uptake, essential for mycobacterial growth, mycobacteria are known to produce Fe^{3+}-specific high-affinity low-molecular-mass compounds, called siderophores, that are able to chelate metal iron from insoluble and protein-bound iron. Two types of siderophores are produced by mycobacteria, the hydrophobic mycobactins and the water soluble carboxymycobactins[23, 24]. Besides siderophores, many pathogens also secrete sphingomyelinases that lyse erythrocytes to gain access to hemoglobin-bound heme. By opposition with most bacterial siderophores, carboxymycobactins are able to remove iron from host transferrin, which role is, alongside with lactoferrin, to transport iron to the cells and control the level of free iron in the blood[24].

Another factor influencing mycobacteria metabolism is the typically low pH. Bactericidal effects of excess of protons are known and may actually be the major mechanism of macrophages to eliminate bacteria. The pH of phagosomes containing *Mtb* can range from pH 6.2 to pH 4.5, depending on whether macrophages are immunologically activated by IFN-γ or not[25]. The pH range in which mycobacteria can grow, like most bacteria, reflect their environment, meaning that saprophytic mycobacteria which reside in soil and water are adapted to a wider pH range than the pathogenic mycobacteria adapted to a specific host. Optimal growth of *Mtb* is observed in enriched media from pH 5.8 to 6.7, almost ceasing replication at pH under 5.5. When high cell densities are considered, the survival range can go down to pH 4.5[25]. Rhode and coworkers (2007) have shown that some polyketide synthases genes and other genes involved in cell wall synthesis, namely *pks2*, *pks3*, *papA3* and *papA1*, are upregulated in *Mtb* CDC1551 after low pH exposure[26].

Even though glycerol is the preferred carbon source used for *in vitro* growth of mycobacteria, some species like *Mb*, *M. africanum* and *M. microti* are unable to grow with glycerol as sole carbon source and pyruvate is added to enable growth. A disruption of glycolysis by a SNP in the gene encoding pyruvate kinase (*pykA*) and the subsequent interruption of the link between the pentose phosphate pathway and pyruvate, that represents the only alternative catabolic pathway available in MTC, makes the introduction of pyruvate in the growth medium mandatory[27].

1.4 The architecture of the mycobacterial cell wall

The mycobacterial cell envelope, a crucial interface with the host, is composed by three major structures (Fig. 1.1). The inner layer resembling a typical peptidoglycan structure, composed by N-acetyl muramic acid (NAM) and N-acetyl glucosamine (NAG), except that, in addition, it presents *Mycobacterium*-specific glycolipids, lipoglycans and lipoproteins. The cell wall core, located outside the plasma membrane, contains peptidoglycan in covalent attachment with arabinogalactan that is esterified to mycolic acids. Mycolic acids are long-chain fatty acids specific of mycobacterial cell wall that are directly related to cell viability and impermeability. This bond between mycolic acids and the arabinogalactan form the inner leaflet of the outer membrane, while the outer layer is composed of several non-covalently attached glycolipids, lipoglycans (such as lipomannan and lipoarabinomannan) and lipoproteins (Fig. 1.2)[28-31]. Amid the *Mycobacterium*-specific components of several pathogenic mycobacteria, namely *Mtb* and *Mb*, are two related families of glycosylated lipids: diphthioceranates (DIP) and phthiocerol dimycocerosate (PDIM) and its variant phenolic glycolipids (PGL)[32]. PGL are known to be associated with impermeability of the cell wall, phagocytosis[33, 34], defense mechanisms against nitrosative and oxidative stress[35] and, theoretically, to the ability of mycobacteria to form biofilms[36].

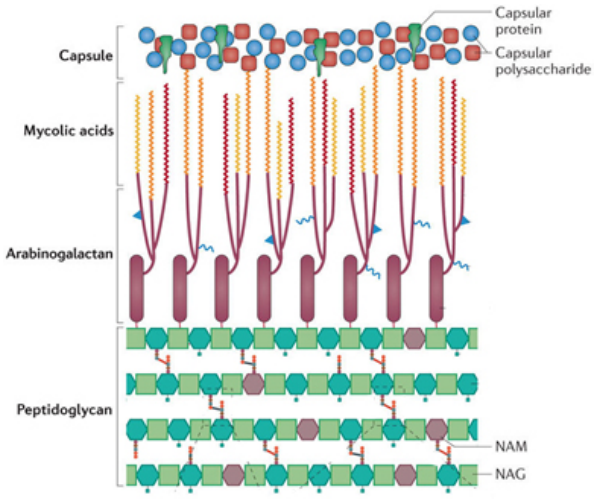

Figure 1.2 – **The architecture of the mycobacterial cell envelope**. Abbreviations: NAM – N-acetyl muramic acid NAG – N-acetyl glucosamine. Adapted from: Kieser *et al.* (2014)[37].

1.4.1 Biosynthetic pathway of PGL production

The genes related to the biosynthesis of PGL belong to the class of polyketide synthases (PKS). There are three types of PKS, classified according to their structure and biosynthetic function. Type I PKS contain multiple catalytic domains and can be classified as modular or iterative. Modular type I PKS have distinct functional domains that are used only once during the formation of the product. On the other hand, iterative PKS have functional domains that intervene repetitively to produce the final polyketide. Type II PKS are composed by several enzymes being

that each one carries a single and distinct catalytic domain that is used iteratively during formation of the polyketide product. Chalcone synthase-like PKS, the type III PKS, represent a more divergent group that, in opposition to types I and II PKS, does not require the involvement of acyl carrier proteins (ACP)[38].

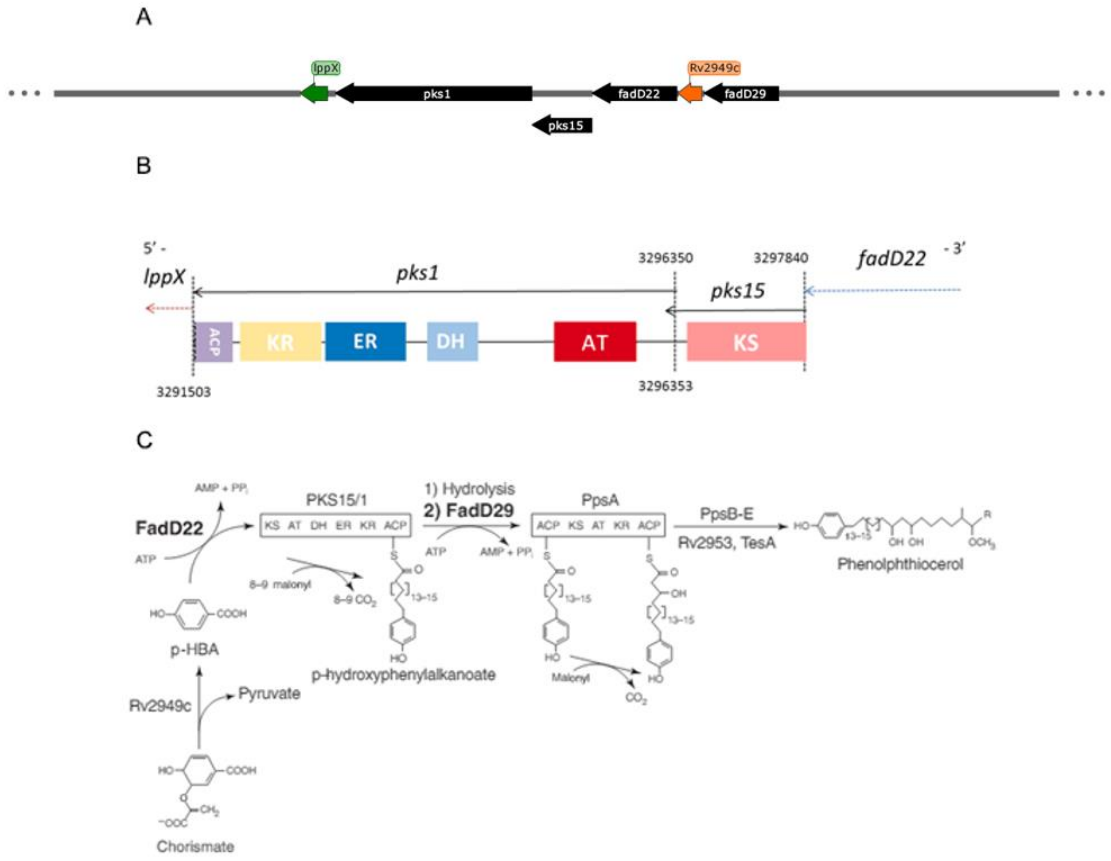

Figure 1.3 – Genomic *locus* of *pks1* and *pks15*, protein domains and their role in the biosynthetic pathway of PGL. A – Schematic representation of the location of pks1 and pks15 in the minus strand of M. tuberculosis H37Rv genome. In black: lipid metabolism. In green: cell wall and cell processes. In orange: intermediary metabolism and respiration. B – Domain organization of Pks1 and Pks15. Adapted from Prata (2016). C -Biosynthetic pathway of phenolphtiocerol moiety of PGL**.** Adapted from Siméone et al. (2010). Abbreviations: KS, ketoacylsynthase; AT, acyltransferase; DH, dehydratase; ER, enoylreductase; KR, ketoreductase; ACP, acylcarrier protein.

Type I PKS are known to be present in Actinobacteria like *Mtb.* Their module is constituted by a minimal set of three domains, namely a ketoacyl synthase (KS) domain, an acyltransferase (AT) domain, and an acylcarrier protein (ACP) domain. This module can also contain all, some or none, of the following domains: keto reductase (KR), dehydratase (DH) and enoyl reductase (ER)[39]. In this group are included two target genes that were explored in this thesis, *pks15* encoding a KS domain and *pks1* encoding KR, DH, ER, AT and ACP domains (Fig. 1.3 B).

The biosynthetic pathway responsible for the synthesis of phenolphthiocerol moiety of PGL comprises several genes encoding type I PKS, namely *ppsA-E* and *pks15*/1 constituting the PDIM + PGL *locus*, which is known to be highly conserved in the PDIM/PGL-producing strains, excepting *M. leprae* in which the *locus* is split into two *loci*[33]. It is genetically proved that *pks15*/1 is involved in PGL production and that it is highly conserved among PGL-producers[40]. Constant and coworkers (2002) documented that *pks1* and *pks15* genes present in PGL-deficient strains, like H37Rv or Erdman, correspond to a single gene *pks15/1* in PGL-producers. This is due to a 7

bp deletion in some *Mtb* strains, and 1 bp deletion in *Mb* strains, generating a frameshift that is responsible for the split in *pks15* and *pks1*[41](Fig. 1.3 A and C).

PGL's phenolphthiocerol moiety production starts with the enzyme encoded by *Rv2949c* that catalyzes the formation of p-hydroxybenzoic acid (p-HBA) that will later be activated by *fadD22*' product, that reveals p-hydroxybenzoyl-AMP ligase activity, and finally elongated with malonyl-CoA as extender unit by *pks15*/1, in a reaction that may comprise 8 to 9 elongation cycles. The product of *fadD29*, a fatty acyl-AMP ligase, will then be responsible for activation of *p*-hydroxyphenylalkanoates, later transferred onto *ppsA'* product and finally elongated with malonyl-CoA and mehtylmalonyl-CoA by PpsB-PpsE to yield the phenolphtiocerol moiety of PGL[42] (Fig. 1.3 C).

1.5 Transcriptional regulation in mycobacteria: the role of σ factors

In *Mtb*, the transcriptional network depends upon the association of the RNA polymerase with one of the 13 σ subunits encoded by *Mtb* genome, namely the essential housekeeping sigma factor (σ^A), the stress-responsive factor (σ^B) or one of the other 11 sigma factors that act as environmental responsive regulators (σ^{C-M}). Those thirteen factors are part of the σ^{70} family, whose members recognize two sequences in the promoter region of their target genes, the -10 element and the -35 element. In *Mtb* promoters, the -10 element is much more conserved than the -35 element [43]. Several studies have been performed to infer the role of each factor and the condition that triggers their activation, initially by analysing expression levels and, more recently, by the construction of deletion strains. The σ^B, σ^F, σ^G, σ^I, and σ^J were shown to be involved in stationary phase regulation[43]. The σ^B subunit, as a major stress response factor, is also involved in regulation during starvation, low and high temperature conditions, hypoxia, oxidative and surface stress[43, 44]. Regulation during starvation also involves the action of σ^D, σ^E and σ^F [43, 44]. In conditions of pH stress, the only factor found to operate was σ^E [43]. For temperature variation, the interference of factors depends on whether the temperature increases or decreases by comparison with the optimal growth temperature. Growth under low temperature stress reveals the involvement of σ^H and σ^I, while under high tempertures it was verified the involvement of σ^H, σ^E and σ^M [43]. When it comes to oxidative stress, interactions of σ^C, σ^E, σ^H and σ^J, have been registered [43, 44]. Under SDS-mediated surface stress, the factors involved were found to be σ^E and σ^H [43, 44]. The presence of this wide variety of sigma factors enables an adaptive transcriptional response for a large set of conditions[43].

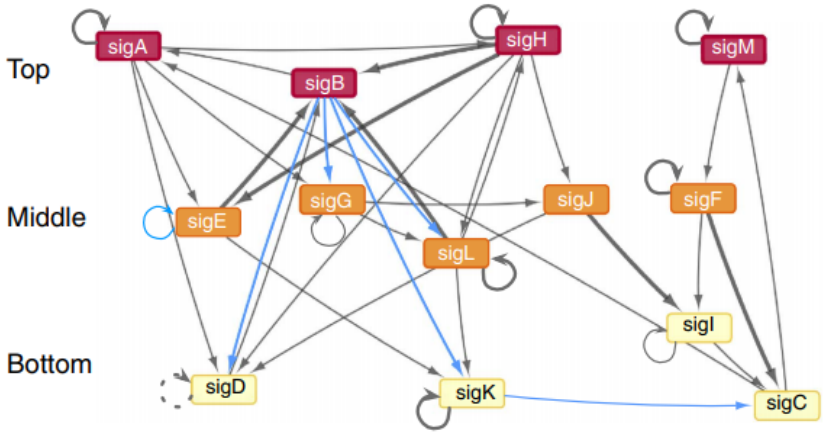

Figure 1.4 - The sigma factor regulatory network of *M. tuberculosis*. Thick lines represent interactions reported in the literature up until 2016. Thin lines represent novel interactions: in blue are those tested and validated in *M. tuberculosis* by Chauhan and coworkers. Adapted from Chauhan *et al*. (2016).

Chauhan and coworkers (2016) performed a reconstruction of the sigma factor regulatory network that allowed to clarify the direct and indirect connections between the 13 factors. This study has defined an hierarchical organization of the factors, as well as the multiple factor usage to respond to specific stresses. Current knowledge advocates a hierarchical organization that comprises three levels: (i) top level: *sigA*, *sigB*, *sigH*, *sigM*; (ii) middle level: *sigE*, *sigF*, *sigG*, *sigJ*, *sigL*; and (iii) bottom level: *sigC*, *sigD*, *sigI*, *sigK* (Fig. 1.4).

1.6 Introduction to the thesis theme and aims

Amongst the cell wall components, PGL have an important role in several aspects of virulence and defense mechanisms. Since *pks1* and *pks15* gene products are critical to defining whether a strain produces PGL or not, it is of major importance to better understand their regulation. Previous studies from our research group have already described the nucleotide and aminoacid diversity of those genes along with their microevolutionary history. The current work had an experimental component and also a bioinformatic framework and was thus divided in two parts: (A) Identification of *pks1* function and transcriptional structure by means of genetics and molecular biology techniques; and (B) *in silico* analysis of *pks1* and *pks15* transcriptional profiles. The thesis aims were thus specifically set to: [A.1] define as to whether or not *pks1* and *pks15* are co-transcribed as a polycistronic transcription unit by cloning the putative promoter sequences on transcriptional fusion vectors and subsequently infer promoter activity by β-galactosidase assays under different growth conditions; [A.2] explain the function of *pks1* through the construction of a transposon-free knock-out mutant based on two parallel methodological approaches relying on phage-mediated mutagenesis and recombineering; and [B.1] identify the regulatory patterns responsible for controlling *pks1* and *pks15* transcription by exploring publicly available large datasets of transcriptome data (RNA-seq), which allowed to define sets of correlated genes according to their gene expression profiles and also outline the transcriptional profile of *pks1* under different stress conditions by means of a differential expression approach.

2. Materials and Methods

2.1 General molecular techniques

The amplification of *pks1*, *pks15* and *fadD22* upstream regions were performed by polymerase chain reaction (PCR) in Biometra Uno II Thermal Cycler or Biometra P Cycler apparatuses, in a final reaction volume of 25 µl [1.25 µl of each primer (Supplementary Table 6.3) at a final concentration of 0.5 µM, 12.5 µl of Phusion® High-Fidelity Master Mix (NEB), 0.75 µl dimethyl sulfoxide (DMSO), 4.25 µl nuclease-free water and 5 µl *Mtb* H37Rv DNA 1:10]. Those reactions were thermocycled at 98ºC for 30 s, followed by 35 cycles of 98ºC for 8 s, 65ºC for 20 s and 72ºC for 30 s, finishing with a final elongation step at 72ºC for 8 min. In order to achieve a higher throughput for some verification tests, colony PCR (cPCR) using pools of five isolates was performed in a final reaction volume of 15 µl [0.75 µl each primer (Supplementary Table 6.3) at a final concentration of 0.5 µM, 7.5 µl NZYTaq 2x Green Master Mix, 1 µl H_2O and 1 µl each isolate DNA 1:100], was thermocycled at 95ºC for 2 min, followed by 35 cycles of 94ºC for 45 s, 59ºC for 45 s and 72ºC for 1 min, finishing with a final elongation step at 72ºC, for 7 min.

Agarose gels were run containing 0.03 µl/ml of GreenSafe Premium (NZYTech) in 1X TBE buffer (89 mM Tris, 89 mM boric acid, and 2 mM EDTA) (Invitrogen) at 90 V. Image acquisition was performed under UV light with Alliance 4.7 system (UVITEC Cambridge).

DNA restriction with the enzymes reported below was performed in a final volume of 30 µl (20 µl DNA, 3 µl reaction Buffer, 1 µl restriction enzyme and 6 µl nuclease-free H_2O), incubated for 1 h at enzyme optimal temperature.

Dephosphorylation of 5'- ends of DNA was performed using Shrimp Alkaline Phosphatase (rSAP) (NEB) in a final reaction volume of 30 µl (20 µl of DNA, 3 µl CutSmart® Buffer (10X), 1.5 µl rSAP and 5.5 µl H_2O), during 1 h at 37ºC. Inactivation was performed by heating at 65ºC, for 5 min.

DNA quantification was performed using the Qubit 2.0 Fluorometer (Invitrogen), with the dsDNA High-Sensitivity Kit, according to manufacturer's instructions using a sample volume of 1 µl.

2.2 Transcriptional analyses

2.2.1 Cloning of regulatory region in transcriptional fusion vectors

The cloning of the putative regulatory regions in transcriptional fusion vectors aims to identify the promoter structure of the selected target genes, as well as to identify which growth conditions trigger those promoters.

Amplification of putative promoter region

For promoter region cloning, primers were designed *in silico* for *pks1*, *pks15* and *fadD22*, targeting roughly 200 bp upstream and downstream the transcription start site (TSS) of each gene. Amplification of the three targets by PCR was performed with Phusion® High-Fidelity DNA Polymerase (NEB). PCR products were visualized in a 1.5% (w/v) agarose gel (NZYTech) and amplicons were excised from gel and purified with QIAquick gel extraction kit (Qiagen), following manufacturer's instructions.

Competent cells preparation

Escherichia coli DH5α and MC1061 were grown overnight (o/n) in 5 ml of Luria-Bertani (LB) medium at 37ºC. Cells were inoculated in 25 ml of LB medium to an initial OD_{600}=0.075 and grown at 37ºC for 3-4 h until OD_{600}=0.4. At this point, cells were maintained on ice for 10 min, followed by centrifugation at 3660 rpm for 10 min at 4ºC. The supernatant was carefully discarded and cells were resuspended in 7.5 ml of an ice cold solution of $MgCl_2$ 80 mM / $CaCl_2$ 20 mM. Cells were pelleted by centrifugation at 3660 rpm for 10 min at 4ºC, supernatant was discarded and cells resuspended in 1 ml of $CaCl_2$ 0.1 M / glycerol 15% for conservation at -80ºC.

Cloning of amplicons

Cloning of amplicons was performed with CloneJET PCR Cloning Kit (ThermoFisher). Ligation reaction included 10 µl of 2X reaction buffer, 25 ng of blunt-end amplicon, 1 µl of pJET1.2/blunt cloning vector, 1 µl of T4 DNA Ligase and H_2O to a final volume of 20 µl, incubated o/n at room temperature (RT). For transformation, an aliquot of *E. coli* DH5α competent cells was used and added to 5 µl of ligation mixture, incubated 30 min on ice, followed by 40 s at 42ºC and 2 min on ice. Nine hundred microliters of SOC medium (20 g/l Casein Peptone, 5 g/l Yeast Extract, 4 g/l Glucose, 10 mM NaCl, 2.5 mM KCl, 5 mM $MgCl_2$ and 5 mM $MgSO_4$) were added to the *E. coli* cells and incubated at 37ºC, 225 rpm, for 1h. Cells were then plated onto LB agar (LA) medium supplemented with ampicillin 100 µg/ml and incubated o/n at 37ºC. Plasmid DNA (pDNA) was extracted from isolated colonies using alkaline lysis method[45] and tested by digestion with *ScaI* (NEB) in order to confirm the correct ligation and excise the desired amplicons. Amplicons were then purified with QIAquick gel extraction kit (Qiagen).

After the first cloning step, the amplicons purified from pJET1.2/blunt were sub-cloned in pSM128 extracted with QIAprep miniprep kit (Qiagen) and digested with *ScaI* (NEB) followed

by dephosphorilation. Ligation of amplicons with pSM128 was performed with Blunt/TA Ligase Master Mix (NEB) in a reaction mixture of 10 µl (5 µl of Blunt/TA Ligase Master Mix and 5 µl of pDNA and amplicon in defined molar ratio). Ligation's molar ratio ranged from 3:1 to 10:1. Transformation of *E. coli* MC1061 was performed with 5 µl of ligation mixture according to the protocol described above. Cells were then plated onto LA medium supplemented with spectinomycin 40 µg/ml and incubated o/n at 37ºC. For DNA extraction from isolated colonies, a loop of a bacterial colony was resuspended in 25 µl of TE, boiled for 10 min and centrifuged at 15000 rpm for 2 min. After centrifugation, DNA from the supernatant was diluted 1:100 in nuclease-free water and tested in pools of five by cPCR. cPCR products were visualized in 1.5% (w/v) agarose gels (NZYTech).

2.3 Construction of knock-out *pks1* mutant strain

The construction of a knock-out *pks1* mutant strain was performed in order to explain the function of *pks1* by comparison of *in vitro* phenotype of the mutant strain to the original phenotype, when mycobacteria are grown in stress conditions. This construction was performed through a two-parallel methodological strategy, including phage-mediated mutagenesis and recombineering.

2.3.1 Specialized Transduction – phage mediated elimination

Specialized transduction of mycobacterial strains contains four stages. The first consists on the construction of an allelic exchange substrate (AES) corresponding to the *pks1* upstream and downstream flanking regions (about 900 bp each) cloned laterally to the hygromycin marker gene present on pYUB854. This step is followed by the construction and packaging of phasmids, meaning that the AES is cloned in the phasmid phAE159 and that construction is subjected to an *in vitro* λ-packaging reaction. After confirmation of the correct structure of the AES in the specialized transducing phage, the next step includes the phasmid transduction in *Mycobacterium smegmatis* (*Msm*) mc^2 155 and the phage lysate preparation. For last, the phage transduction of *Mtb* is performed using the previously prepared phage lysate. In this work, *E. coli* HB101 containing *pks1*/pYUB854 and phAE159 were grown in LB medium supplemented with hygromycin 50 µg/ml and ampicillin 100 µg/ml, respectively. Cosmid DNA was extracted using NZYMiniprep (NZYTech) and phasmid DNA using QIAprep miniprep kit (Qiagen). DNA extraction was followed by enzymatic digestion with 1 µl *PacI* (NEB), 2 µl CutSmart Buffer (NEB) for 2 h at 37ºC, and posterior inactivation (10 min, 65ºC). An 0.8% (w/v) agarose gel (NZYTech) of both digests was run for 1 h 30 min at 90 V. Gel purification was done using Qiagen QIAEX II kit for phAE159 and Qiagen QIAquick gel extraction kit for *pks1*/pYUB854. Following, *pks1*/pYUB854 was dephosphorylated and purified with QIAquick gel extraction kit (Qiagen). Ligation reaction was performed with purified 1000 ng of phAE59 and 250 ng of *pks1*/pYUB854 (2 h at RT followed by 4°C o/n). At the same time, *E. coli* HB101 was grown o/n in 5 ml of LB+10 mM $MgSO_4.7H_2$+0.2% maltose, 37°C and 100 rpm. The day after, 25 ml of the same medium was inoculated with the o/n culture with initial OD_{600} of 0.05 until it reached 0.6/0.8 (37°C, 100 rpm). After growth, 1 ml of culture was centrifuged during 3 min at 13000 rpm and pellet was ressuspended in 400 µl of MP buffer (50 mM Tris/HCl, 150 mM NaCl, 10 mM $MgCl_2$, 2 mM $CaCl_2$). *In vitro* packaging was performed with 5 µl of o/n ligation reaction placed on top of Gigapack III XL (Agilent Technologies), incubated 2 h at RT, and 200 µl of SM buffer (5.8 g NaCl, 2 g $MgSO_4.7H_2O$, 50 ml 1M Tris/HCl, 0.5 ml agar 2%, 94.5 ml H_2O) was added. The 1.5 ml microtube was centrifuged for 40 s, placing 200 µl of the supernatant in a new 1.5 ml microtube, remaining 50 µl in the Gigapack 1.5 ml microtube. To both tubes, 200 µl of *E. coli* HB101 were added, following static incubation at 30ºC and addition of 750 µl of LB. The tubes were then incubated at 37ºC, 1 h at 100 rpm and after briefly centrifuged – 900 µl of supernatant

and the remaining 100 µl were inoculated into LA medium supplemented with hygromycin 150 µg/ml and incubated at 37ºC until colonies were visible. Each colony was replated and grown in LB medium supplemented with hygromycin 50 µg/ml, pDNA was extracted, digested with *PacI* for 2 h and a 0.8% gel run was performed to check for transformants. pDNA extraction was carried out with QIAprep miniprep kit (Qiagen), Large Construct Kit (Qiagen) or JETSTAR 2.0 Plasmid Miniprep Kit (Genomed).

2.3.2 Mycobacterial Recombineering – knock-out by homologous recombination

Bacterial strains and growth media

Msm mc^2155 was inoculated into Middlebrook 7H9 (BDDifco™) or Middlebrook 7H10 (BDDifco™), both supplemented with ADS 10% (v/v) (dextrose 20 g/l, NaCl 8.5 g/l and bovine albumin fraction V 50 g/l) and 2 ml/l of glycerol 86% and incubated at 37ºC with agitation (200 rpm) during 1-2 days. For preparation of recombineering strain, *Msm* mc^2155 was inoculated into 7H9 induction medium (Middlebrook 7H9 (BDDifco™), ADS 10% (v/v), 2 ml/l of glycerol 86% and 10 ml/l 20% succinate).

Electrocompetent cells preparation

Msm mc^2155 cells were grown in 50 ml of 7H9 medium until OD_{600}= 0.8. Cells were then transferred to centrifuge tubes and pelleted by centrifuging at 4230 rpm for 10 min. After centrifugation, the supernatant was carefully discarded, and cells were washed five times with 10% ice-cold glycerol with 1/2, 1/4, 1/8, 1/10 and 1/25 of initial culture volume. After washed, cells were aliquoted and stored at -80ºC. Both labware and solutions were maintained at 4ºC.

Preparation of recombineering strain

An aliquot of *Msm* mc^2155 electrocompetent cells was incubated on ice for 10 min, followed by addition of 50 ng of pJV53 DNA and incubation on ice for more 10 min. The mixture was transferred into a pre-chilled electroporation 0.2 mm cuvette and electroporated with the following conditions: 2.5 kV, 1000 Ω, 25 µF in Gene Pulser® II Eletroporation System (Biorad). Cells were then inoculated on 7H9 medium for recovery and incubated for 2 h at 37ºC. This step was followed by inoculation of the cells into 7H10 medium supplemented with kanamycin 20 µg/ml and incubated for 3 days at 37ºC. An isolated colony, whose confirmation was made by digestion with *XbaI* and *SpeI*, was then inoculated in 7H9 induction medium and grown with o/n shaking at 37°C. Once the cells reached OD_{600} = 0.5, acetamide was added to a final concentration of 0.2%. The culture was grown under shaking at 37°C, for 3 h. Electrocompetent cells were then prepared as described above.

Transformation of the recombineering strain with allelic exchange substrate

Newly prepared cells were then transfected with 100 ng of AES according to the same conditions described above. Cells were recovered by incubation on 7H9 medium, for 4 h at 37ºC. After recovery, those cells were plated onto 7H10 medium, supplemented with kanamycin 20 µg/ml and hygromycin 50 µg/ml, and incubated for two days at 37ºC.

2.4 *In silico* analysis of regulatory data of selected genes

Regulatory information of *pks1* was gathered from international databases such as: Tuberculist[46], National Center for Biotechnology Information (NCBI)[47], TB Database[48, 49] and MTB Network Portal[50] (visited from 10/2016 to 01/2017). Location of putative ribosomal binding sites (RBS) was inferred for *pks1* with *Prokaryotic Dynamic Programming Genefinding Algorithm*

(PRODIGAL)[51]. Synteny analyses were performed for *pks1*, *pks15* and *fadD22* using SyntTax (Prokaryotic Synteny & Taxonomy Explorer)[52].

2.4.1 RNA-seq data analysis

For expression analyses, 77 experiments were collected from NCBI database, 60 from *Mtb* strains (*Mtb* H37Rv, *Mtb* CDC1551 and *Mtb* clinical isolates) and 17 from *Mb* strains (*Mb* clinical isolates and *Mb* BCG str. Pasteur 1173P2), constituting a set of 25 experimental conditions in five datasets for *Mtb* and six experimental conditions in two datasets for *Mb* (Tables 2.1 and 2.2). This analysis was performed for a set of 52 genes for *Mtb* and 50 genes for *Mb*, including *pks1*, *pks15*, *fadD22*, *fadD29* genes included in bicluster modules 0211 and 0490 from MTB Network Portal that represent co-regulated genes, and genes encoding PKS and σ factors (Table 2.3). For each experiment, reads were extracted in FASTQ format using NCBI SRA Toolkit[53]. Those FASTQ files were mapped against a reference genome, *Mtb* H37Rv (RefSeq code: NC_000962.3) or *Mb* AF2122/97 (RefSeq code: NC_002945.3) with TopHat v.2.1.0.[54, 55], using default settings to produce a BAM file containing a list of read alignments. Transcript identification and counting was later performed with bias correction by Cufflinks v.2.2.1.0[54, 56] using as reference the annotation of the genomes listed above. Cufflinks was used to calculate the relative abundance of each gene in Reads Per Kilobase of exon model per Million mapped reads (FPKM). FPKM values were transformed by $\log_{10}$, heatmaps were plotted by ClustVis[57] and dendrograms were computed using BioNumerics v4.0 (Applied Maths) calculating Pearson correlation coefficient and the unweighted pair group method with arithmetic means (UPGMA) as the agglomerative clustering algorithm. Pearson correlation coefficient was calculated using GraphPad Prism 6 and correlation network was plotted using Cytoscape[58, 59]. For differential expression analysis, htseq-count[60] was used to count reads mapped to each gene and DESeq2[61] was used to determine differentially expressed genes from count tables using Wald statistic test with p-value adjusted for multiple testing with the Benjamini-Hochberg procedure (α=0.05). Data was plotted as bar plots using GraphPad Prism 6. The fluxogram underlying these analyses is represented in Fig. 2.1. The public server at usegalaxy.org[62] was used to analyse the data with NCBI SRA Toolkit, TopHat, Cufflinks, htseq-count and DESeq2.

[46] Tuberculist available at https://mycobrowser.epfl.ch/; previously available at https://tuberculist.epfl.ch/

[47] NCBI available at https://www.ncbi.nlm.nih.gov/

[48, 49] TB Database previously available at https://www.tbdb.org

[50] MTB Network Portal available at http://networks.systemsbiology.net/mtb/

[51] PRODIGAL source code available at https://github.com/hyattpd/Prodigal; previously available at http://compbio.ornl.gov/prodigal/ (visited on 11/05/2017)

[52] SyntTax available at http://archaea.u-psud.fr/synttax/ (visited on 03/12/2016)

[53] NCBI SRA Toolkit source code available at https://www.ncbi.nlm.nih.gov/sra/docs/toolkitsoft/ (visited on 17/05/2017)

[54, 55] TopHat source code available at http://ccb.jhu.edu/software/tophat/index.shtml (visited on 17/05/2017)

[54, 56] Cufflinks source code available at https://github.com/cole-trapnell-lab/cufflinks (visited on 17/05/2017)

[57] ClustVis available at http://biit.cs.ut.ee/clustvis/ (visited on 01/09/2017)

[58, 59] Cytoscape available at http://www.cytoscape.org/ (visited on 03/11/2017)

[60] htseq-count available at https://htseq.readthedocs.io/en/release_0.9.1/ (visited on 05/08/2017)

[61] DESeq2 available at http://bioconductor.org/packages/release/bioc/html/DESeq2.html (visited on 05/08/2017)

Table 2.1 – List of SRA codes used for transcriptome analysis with the corresponding code used in the current work and the experiment brief description for *Mtb* strains.

Code	Description	SRA
H37Rv-Log	*Mtb* H37Rv strain in log phase	SRR1822433 SRR1822434 SRR1822435
H37Rv-Early	*Mtb* H37Rv strain in early dormancy phase (K^+-deficient medium for 14 days)	SRR1822436 SRR1822437 SRR1822438
H37Rv-Medium	*Mtb* H37Rv strain in medium dormancy phase (K^+-deficient medium; 10 days after addition of rifampicin)	SRR1822439 SRR1822440 SRR1822441
H37Rv-Late	*Mtb* H37Rv strain in late dormancy phase (K^+-deficient medium; 20 days after addition of rifampicin)	SRR1822442 SRR1822443 SRR1822444
CDC-Gly-pH 5.7	*Mtb* CDC1551 strain grown in glycerol at pH 5.7	SRR1023807 SRR1023808
CDC-Gly-pH 7	*Mtb* CDC1551 strain grown in glycerol at pH 7	SRR1023805 SRR1023806
CDC-Pyr-pH 5.7	*Mtb* CDC1551 strain grown in pyruvate at pH 5.7	SRR1023811 SRR1023812
CDC-Pyr-pH 7	*Mtb* CDC1551 strain grown in pyruvate at pH 7	SRR1023809 SRR1023810
H37RV-DE	*Mtb* H37Rv strain grown in dextrose at exponential phase	SRR896652
H37RV-DS	*Mtb* H37Rv strain grown in dextrose at stationary phase	SRR896653
H37RV-FE	*Mtb* H37Rv strain grown in long fatty acids at exponential phase	SRR896654
H37RV-FS	*Mtb* H37Rv strain grown in long fatty acids at stationary phase	SRR896655
H37Rv-Glu	*Mtb* H37Rv strain grown in 0.4% glucose	SRR1917700 SRR1917701 SRR1917702
H37Rv-LI-1day	*Mtb* H37Rv strain grown in low iron for 1 day (chelated Sauton's medium with 100 μM 2,2'-bipyridine)	SRR1917706 SRR1917707 SRR1917708
H37Rv-LI-1week	*Mtb* H37Rv strain grown in low iron for 1 week (chelated Sauton's medium with 100 μM 2,2'-bipyridine)	SRR1917709 SRR1917710 SRR1917711
H37Rv-HI	*Mtb* H37Rv strain grown in high iron (with 150 μM $FeCl_3$)	SRR1917703 SRR1917704 SRR1917705
HN878	*Mtb* HN878 strain grown under regular *in vitro* conditions	SRR1917716 SRR1917717 SRR1917718
H37Rv-aera-1day	*Mtb* H37Rv strain 1 day after reaeration	SRR3725588 SRR3725589

		SRR3725590
H37Rv-aera-2day	*Mtb* H37Rv strain 2 day after reaeration	SRR3725591 SRR3725592 SRR3725593
H37Rv-aera-3day	*Mtb* H37Rv strain 3 day after reaeration	SRR3725594 SRR3725595 SRR3725596
H37Rv-aera-4day	*Mtb* H37Rv strain 4 day after reaeration	SRR3725597 SRR3725598 SRR3725599
H37Rv-hypoxia	*Mtb* H37Rv strain grown in hypoxia	SRR3725585 SRR3725586 SRR3725587
Extract N0031	*Mtb* clinical strain N0031	ERR219205 ERR219199
Extract N0145	*Mtb* clinical strain N0145	ERR219204 ERR219198
Extract N0153	*Mtb* clinical strain N0153	ERR219194 ERR219201

Table 2.2 – List of SRA codes used for transcriptome analysis with the corresponding code used in the current work and the experiment brief description for *Mb* strains.

Code	**Description**	**SRA**
Mb-Log	*Mb* BCG str. Pasteur 1173P2 grown to log phase	SRR1915476 SRR1915477 SRR1915478
Mb-Starv-4day	*Mb* BCG str. Pasteur 1173P2 grown for 4 days in starvation	SRR1915479 SRR1915480 SRR1915481
Mb-Starv-10day	*Mb* BCG str. Pasteur 1173P2 grown for 10 days in starvation	SRR1915482 SRR1915483 SRR1915484
Mb-Starv-20day	*Mb* BCG str. Pasteur 1173P2 grown for 20 days in starvation	SRR1915485 SRR1915486 SRR1915487
Mb-Resuscitation	*Mb* BCG str. Pasteur 1173P2 after starvation	SRR1915488 SRR1915489 SRR1915490
Mb	*Mb* clinical strain grown under regular *in vitro* conditions	SRR1020650 SRR1020651

Table 2.3 – List of genes under expression analysis in both *Mtb* and *Mb* with group classification.

	Mtb	***Mb***
Genes of interest	*lppX**	*lppX*
	*pks1***	*pks15/1*
	*pks15***	
	*fadD22***	*fadD22*
	*Rv2949c**	*Mb2973c*
	*fadD29**	*fadD29*
Genes encoding polyketide synthases	*pks2*	*pks2*
	pks3	*pks3*
	pks4	
	pks5	*pks5*
	pks6	*pks6a*
		pks6b
	pks7	*pks7*
	pks8	*pks8*
	pks9	*pks9*
	pks10	*pks10*
	pks11	*pks11*
	pks12	*pks12*
	pks13	*pks13*
	pks16	*pks16*
	pks17	*pks17*
Members of bicluster module 0211	*whiB5*	*whiB5*
	Rv0024	*Mb0024*
	Rv0268c	*Mb0274c*
	rpsN2	*rpsN2*
	rpsR2	*rpsR2*
	rpmG1	*rpmG1*
	rpmB2	*rpmB2*

	Mtb	***Mb***
Members of bicluster module 0490	*Rv0421c*	*Mb0429c*
	lpqT	*lpqT*
	lppR	*lppR*
	otsB2	*otsB2*
	emrB	*emrB*
	lpqU	*lpqU*
	lepA	*lepA*
	Rv3300c	*Mb3328c*
	echA18	*echA18*
	amiD	*amiD*
	PPE67	
	aftA	*Mb3821*
Genes encoding sigma factors	*sigA*	*sigA*
	sigB	*sigB*
	sigC	*sigC*
	sigD	*sigD*
	sigE	*sigE*
	sigF	*sigF*
	sigG	*sigG*
	sigH	*sigH*
	sigI	*sigI*
	sigJ	*sigJ*
	sigK	*sigK*
	sigL	*sigL*
	sigM	*sigMa*
		sigMb

* Member of bicluster module 0490

** Member of bicluster module 0211 and module 0490

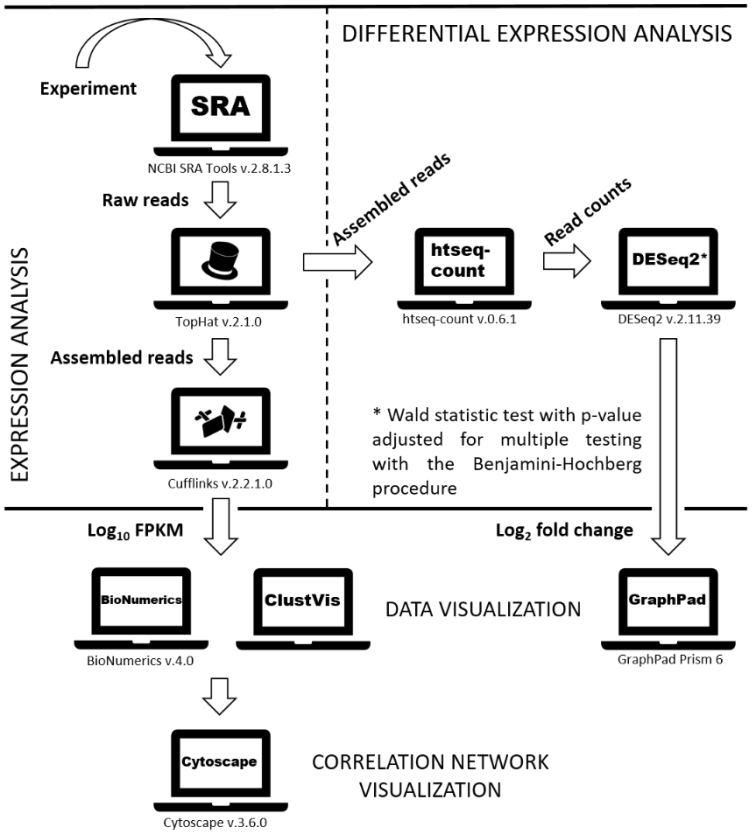

Figure 2.1– Schematic representation of the *in silico* analysis of RNA-seq data performed in the current work. Expression analysis was performed by an adapted Tuxedo pipeline including TopHat and Cufflinks. Differential expression analysis was performed by generation of count tables by htseq-count with TopHat output and introduction of count tables as input for analysis by DESeq2. Data visualization was performed in the form of a heatmap with ClustVis and as dendrograms using Bionumerics for expression data and in the form of bar plots with GraphPad Prism 6 for differential expression data. Expression data was then analysed as correlation networks produced by Cytoscape.

3. Results and Discussion

3.1 Putative promoter region cloning in pSM128

Cloning of the corresponding upstream region was sought as a key step to identify the promoter and operon structure of *pks1*, *pks15* and *fadD22*, as well as to clarify which stress conditions trigger their transcription. pSM128 is a transcriptional cloning shuttle vector that replicates both in *E. coli* and mycobacteria and contains a unique *ScaI* site before *lacZ*, allowing blue/white screening, and an antibiotic resistance cassette, enabling selection by streptomycin and spectinomycin. This method includes the cloning of a putative regulatory region on the upstream region of *lacZ*[63]. Once the ligation reaction is performed, *E. coli* MC1061 is transformed enabling blue/white screening, since this strain presents a ΔlacX74 mutation, meaning that the *lac* operon is almost completely deleted, with exception of *lacA*[64] (Fig. 3.1).

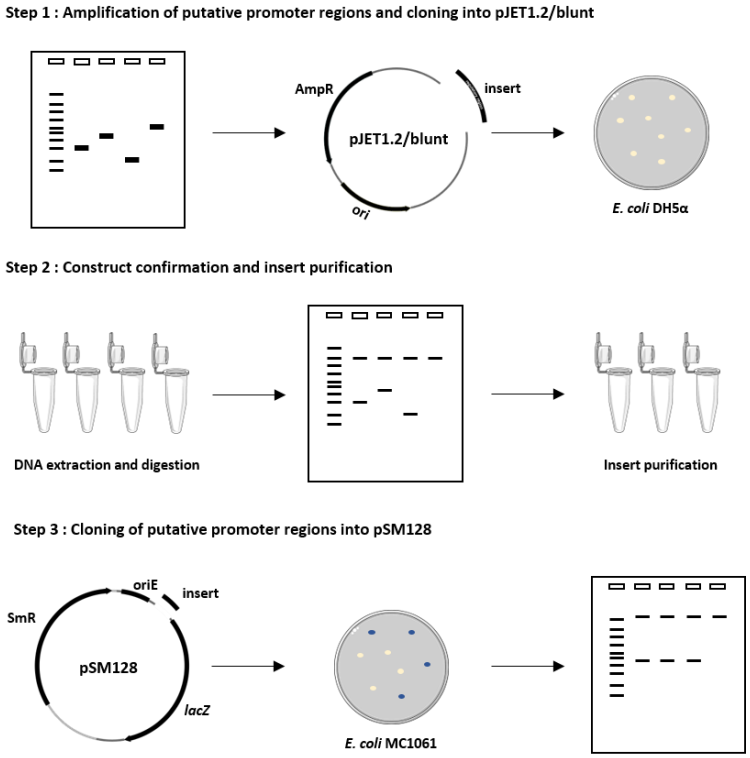

Figure 3.1 – Schematic representation of putative promoter region cloning. Step 1 comprises the amplification by PCR of the putative promoter regions of the selected genes of interest, cloning into pJET1.2/blunt and transformation of *E. coli* DH5α. Step 2 comprises the DNA extraction and digestion with *ScaI* for construct confirmation and insert purification. Step 3 comprises subcloning of the previously purified inserts into pSM128, transformation of *E. coli* MC1061 for blue/white screening and construct confirmation by cPCR.

In this work, we first amplified the upstream regions of *pks1* (409 bp), *pks15* (508 bp) and *fadD22* (336 bp) possibly encompassing the regulatory region of these genes, those amplicons were purified, and cloning was performed in pJET1.2/blunt, followed by construct confirmation (Figs. 3.2 and 3.3).

Figure 3.2 – Schematic representation of primer hybridization for putative promoter region amplification. The relative location of primers from supplementary Table 6.3 against *Mtb* H37Rv genome is represented. In black: selected genes of interest; in green: putative ribosome binding site (RBS) for *fadD22* and *pks1*; and in grey: amplified regions for *fadD22*, *pks15* and *pks1*.

Following cloning in pJET1.2/blunt, we tried to subclone each of the putative regulatory regions in the shuttle pSM128, with a 3:1 vector-insert molar ratio and using 50 ng of vector. As none of the colonies resulting from this transformation presented the correct construct, the cloning was then attempted with higher vector-insert ratios, namely, 5:1 and 7:1. Once again, those ratios were proven to be inefficient for the desired construct and ligation was tried once with 10:1 ratio. For each of those ratios, ligation was tried in two different conditions, first ligation was set for 2 h at RT and o/n at 4ºC and latter we tried to set it o/n at 16ºC. For ligation improvement, it was also tried the addition of polyethyleneglycol (PEG) to promote DNA molecules cohesion. As none of the strategies used led us to a positive transformant, ligations were then performed with a blunt

cloning specific ligase with 7:1 ratio. Even though our best efforts were put in this method, it was not possible to get the correct construct in the shuttle vector to follow this protocol.

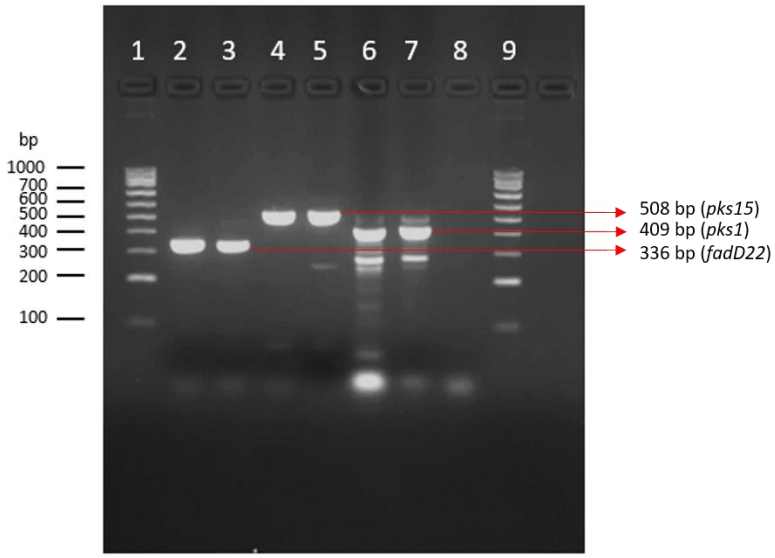

Figure 3.3 – Gel electrophoresis of putative regulatory regions of *fadD22*, *pks15* and *pks1* amplified by PCR. Lanes 1 and 9: NZY Ladder V (molecular marker); lanes 2-3: amplicon obtained with fadD22_F and fadD22_R (336 bp); lanes 4-5: amplicon obtained with *pks15*_F and *pks15*_R (508 bp); and lanes 6-7: amplicon obtained with *pks1*_F and *pks1*_R (409 bp).

3.2 Construction of *Mtb pks1* knock-out mutant

Construction of knock-out mutants remains a hard procedure due to inefficient DNA uptake and very low recombination yield in mycobacteria[65], increased by the high proportion of illegitimate recombination in mycobacteria[66].

3.2.1 Phage-mediated mutagenesis

The construction of a knock-out *pks1* mutant strain was initiated through phage-mediated mutagenesis. This mutagenesis is based on a specialized transduction protocol that allows directed mutagenesis using an allelic exchange substrate (AES) to disrupt the target gene. Conceptually, this method implies the preparation of a *pks1*::hygromycin allele. For that, *pks1* flanking regions (≈ 1 kb) should be cloned into cosmid pYUB854 (hyg^R) and then into the conditionally replicating (temperature-sensitive) shuttle vector phAE159 (Amp^R) to generate a specialized transducing mycobacteriophage. This shuttle vector feature allows replication of mycobacteriophage genome at the permissive temperature of 30ºC but prevents replication at nonpermissive temperature of 37ºC. Transduction is performed at nonpermissive temperature which supposedly results in higher delivery efficiency of the recombination substrate. Flanking the antibiotic-resistance gene that permits selection of transductants with AES are resolvase recognition sites to allow the use of plasmid-encoding *tnpR* to excise the resistance gene and generate an unmarked mutant[67] (Fig. 3.4).

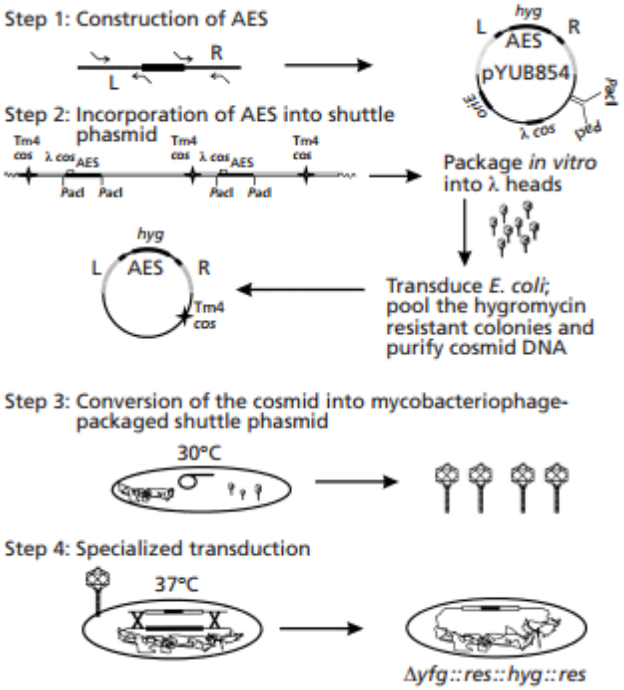

Figure 3.4 – Schematic representation of phage-mediated mutagenesis. Step 1 comprises the construction of AES into the cosmid pYUB854. Step 2 comprises the cloning of AES into shuttle phasmid phAE159, followed by *in vitro* packaging into λ heads and transduction of *E. coli* HB101. Step 3 comprises the conversion of cosmid into mycobacteriophage-package shuttle phasmid using *Msm* mc^2 155 grown at non-permissive temperature. Step 4 comprises the specialized transduction of *Mtb* strains using phage lysates.

In this work, we attempted several times the ligation of a previously confirmed pYUB854_*pks1* (5.7 kb) construction with phasmid phAE159 (47 kb) in order to obtain a shuttle vector for electroporation into *E. coli* HB101 and *Msm* mc^2 155 (Figure 3.5). After the ligation reaction, the transformation of *E. coli* HB101 was attempted and the colonies obtained were grown in hygromycin supplemented medium. The pDNA from those isolates was extracted using different extraction protocols and commercial systems, but due to low yield it was not possible to confirm the hypothetic transformation by digestion. Since the full construct would complete approximately 53 kb, most pDNA extraction kits are not suitable to this plasmid size. Further work will include new approaches to complete the extraction with better yield in order to confirm transformation by digestion.

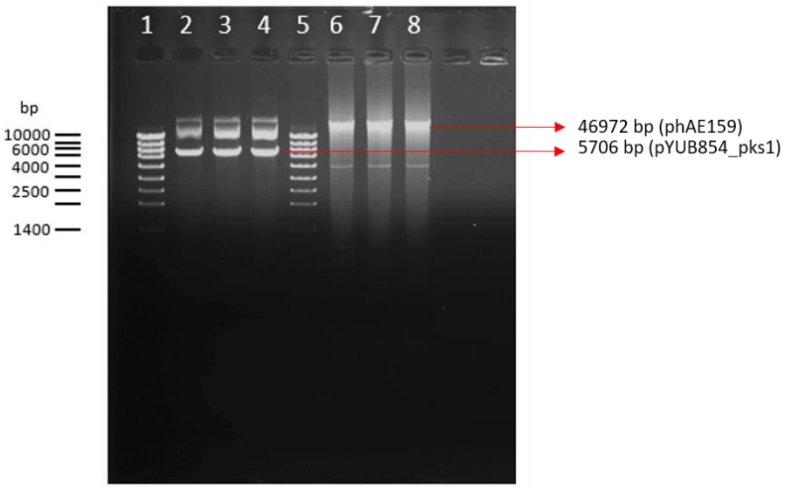

Figure 3.5 – Gel electrophoresis of pYUB854_*pks1* and phAE159. Lanes 1 and 5: NZYDNA Ladder II (molecular marker); Lanes 2-4: pYUB854_*pks1* digested with *PacI*; and Lanes 6-8: phAE159 digested with *PacI*.

3.2.2 Mycobacterial Recombineering

Recombineering constituted the alternative approach to the specialized transduction method for genetic manipulation of mycobacteria aiming the construction of a *pks1* knock-out mutant. This approach was based on a method developed in *Escherichia coli* using phage-encoded recombination proteins. The mycobacterial homologs of those proteins are found only on mycobacteriophage Che9c encoded by genes 60 and 61[65, 68]. For the present method, pJV53 plasmid contains an inducible acetamidase promoter controlling the *gp60* and *gp61* coding sequences. The first step of this protocol includes the electroporation of *Msm* mc^2 155 with pJV53. After induction of acetamidase promotor, there is a second electroporation event for insertion of the AES. Also, as described for specialized transduction, this AES includes resolvase recognition sites that allow posterior excision of antibiotic-resistance gene to generate an unmarked mutant[68](Fig. 3.6).

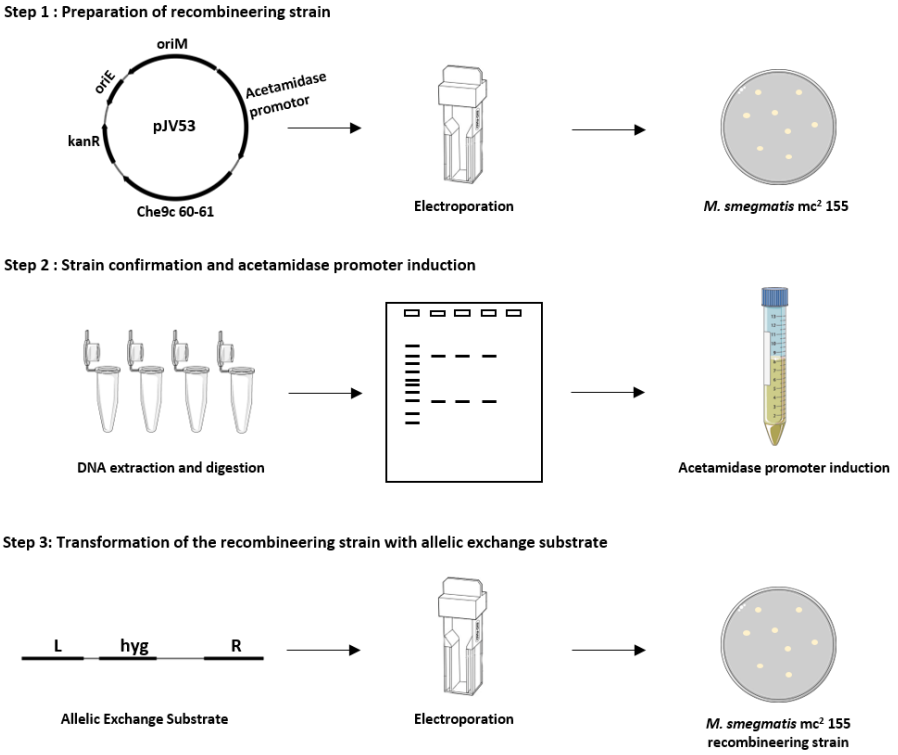

Figure 3.6 – Schematic representation of homologous recombination by mycobacterial recombineering. Step 1 comprises the preparation of a *Msm* mc^2 155 recombineering strain by electroporation of pJV53. Step 2 comprises the DNA extraction and digestion with *XbaI* and *SpeI* for strain confirmation and acetamidase promoter induction in the recombineering strain. Step 3 comprises the electroporation of AES into *Msm* mc^2 155 recombineering strain.

In this work, electroporation of pJV53 circular DNA was performed leading to four isolates confirmed by digestion. Those four isolates were subjected on several occasions to induction by acetamide and transformed by electroporation with AES (Fig. 3.7). Even though optimal DNA amounts were used in both electroporations, circular DNA is less susceptible to exonucleases and therefore more persistent than linear DNA, meaning that electroporation with AES is more likely to be unsuccessful. Future work will include optimization of electroporation with AES in order to produce some isolates to test.

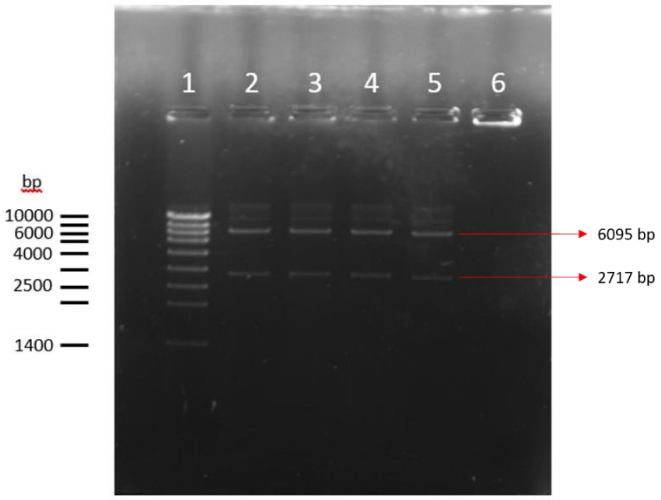

Figure 3.7 – Gel electrophoresis of pJV53 extracted from 4 isolates of *Msm* mc^2 155. Lane 1: NZYDNA Ladder II (molecular marker); Lanes 2-5: pJV53 from 4 isolates digested with *SpeI* and *XbaI* (6 kb + 2.7 kb)

3.3 Regulatory pattern of selected genes of interest

3.3.1 Organization of the genetic *locus* and predicted regulatory data

Mtb pks1 and *pks15* (*pks15/1* in *Mb*) together encode a polyketide synthase with six identified domains involved in the synthesis of PGL. Encoded in the minus strand, from position 3291503 to 3296353 for *pks1* and, from position 3296350 to 3297840, for *pks15*, it is shown to have a functional cooperation with *fadD22*, a bidomain initiation module. Regulatory information was gathered across TB Database[48, 49], Tuberculist[46] and MTB Network Portal[50]. MTB Network Portal[50] reports information produced by cMonkey[69] algorithm that demonstrates that *pks1* and *pks15* belong to the same two biclusters. Biclusters are sets of co-regulated genes defined by cMonkeyaccording to mRNA-based expression levels, *de novo* identification of transcription factor binding motifs and pre-established association pathways[69]. The *pks1* and *pks15* genes are placed in bicluster modules 0211 and 0490, with residual values of 0.5 and 0.57, respectively, meaning that bicluster module 0211 presents a tighter expression profile amongst its members which should also indicate better bicluster quality and thus more certainty associated with co-expression. Also related are the two genes upstream of *pks15*, *fadD22*, included in the same two modules, and *Rv2949c*, included in module 0490. MTB Network Portal[50] also indicates that both *pks1* and *pks15* are regulated by the products of seven genes (positively by *Rv0042c*, *sigK*, *Rv2258c* and *Rv3557c*; negatively by *sigB*, *Rv2745c* and *Rv3583c*). Furthermore, according to chIP-seq data, *pks1* is bound by the transcription factor *Rv3830c* with no differential expression reported. Operon structure is undetermined: TB Database[48, 49] suggests four different combinations of six genes (*fadD29*, *Rv2949c*, *fadD22*, *pks15*, *pks1* and *lppX*), while MTB Network Portal[50] suggests an operon composed by five genes (*fadD29*, *Rv2949c*, *fadD22*, *pks15* and *pks1*). All those genes are involved in the synthesis of phenolphthiocerol moiety of PGL, except *lppX* whose function is unknown. Besides that, *pks1* was previously identified in the cell membrane fraction of *Mtb* H37Rv using liquid chromatography/mass spectrometry[70] and stated as non-essential for *in vitro* growth of *Mtb* H37Rv by transposon mutagenesis[71]. The RBS prediction for *pks1* resulted in a GGGGG sequence (3291486-3291491) with a spacer region of 12 bp between RBS and TSS.

A neighbouring gene comparison was done to evaluate and compare the conservation of *pks1 locus* in *Mtb* genome using SyntTax with Pks1 sequence from *Mtb* H37Rv as query. Among the 210 *Mtb* accession codes available at SyntTax, for *pks15* and *fadD22*, the percentages with synteny score above 80 were 90.5% and 99.5%, respectively. For *pks1*, 89.5% of the accession codes used for analysis presented a score higher than 80 (Fig. 3.8). When analysing local genomic

conservation, it was possible to identify that homology to Pks1 exists in some genomes with Pks15/1 and in others with Pks1. Also, it was showed that Rv2949c, FadD22, LppX, Rv2944 and Rv2943 present a regular pattern along almost every genome analysed, while when analysing Pks1, it presents an irregular pattern in genomes with score under 80. Only by comparing the top three scored genomes, it was also possible to identify that, for *Mtb* accession code 0B070XDR aa7360151, the element upstream of Rv2949c does not present homology with FadD29 and that homology was found for a protein encoded in the opposite strand of *pks1*.

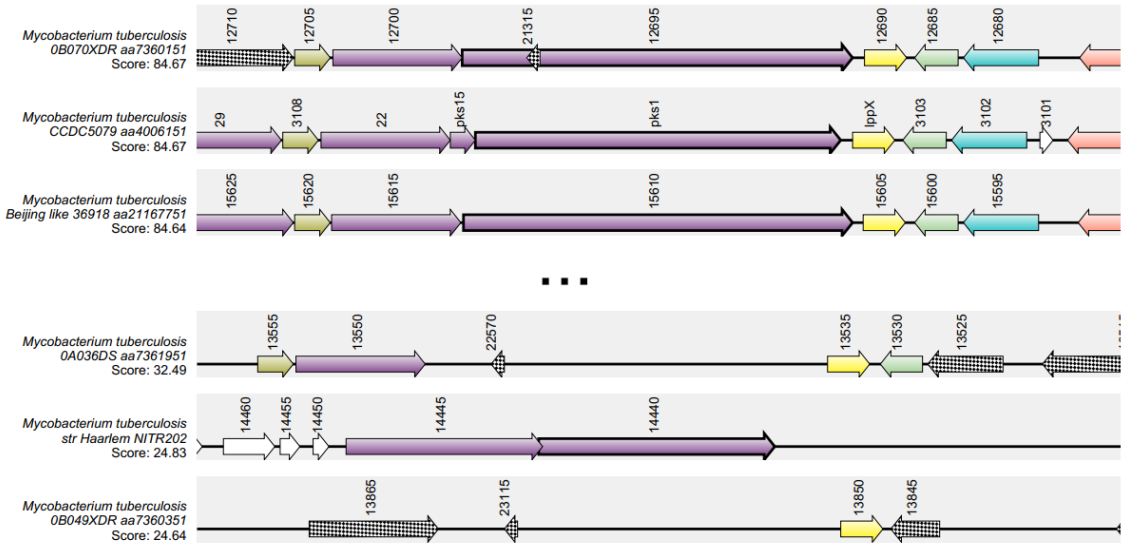

Figure 3.8 – Representation of top three and bottom three scores from synteny analysis. Top 3 and bottom 3 synteny scores for *pks1* as predicted by SyntTax. Color code represents matching proteins across genomes.

3.3.2 Correlation analyses of *pks1* and *pks15* with genes encoding polyketide synthases and σ factors

RNA-seq is a methodological approach for transcriptome profiling that takes advantage of deep-sequencing technologies to get a more precise transcript measurement at the whole genome level, that was used here to infer the expression profile of the defined genes of interest.

Data gathered was analysed by alignment against a reference genome, read counting and calculation of FPKM, a proxy for gene expression in each condition. FPKM is a relative value, meaning that it varies according not only to read count, but also to the total number of reads obtained for each experiment. It is thus not possible to compare directly the expression across different conditions. In contrast, we performed direct comparison amongst genes (Table 2.3) within the same experiment since there is no missing data and thus variation will be equal for all. Those FPKM values were transformed by $\log_{10}$ and plotted in a heatmap (Fig. 3.9). To understand the similarity between FPKM profiles, a dendrogram was generated, using 85% similarity as a cut-off for cluster formation. This cut-off was established to include the top quarter of similarity values in this analysis. We then obtained a total of 20 clusters, being 14 single member clusters (Fig. 3.9). Among the remaining four clusters, we found cluster IV, including genes belonging to the putative polycistronic structure (*pks1*, *fadD22*, *Rv2949c* and *fadD29*), and also some of *pks1* paralogs (*pks6*, *pks12* and *pks9*). The cluster III is also relevant, once it is constituted by three genes encoding polyketide synthases, namely *pks4*, *pks3* and *pks2*. The cluster VI contains the larger number of members comprising members of bicluster modules 0211 and 0490 and genes encoding σ factors and PKS genes. To obtain an approach more focused in direct relation between genes, we constructed a correlation network with pairs of genes exhibiting correlation factors above 0.9 (Figure 3.10). The correlation values between *pks1*, *fadD22*, *Rv2949c* and *fadD29* are all above 0.9.

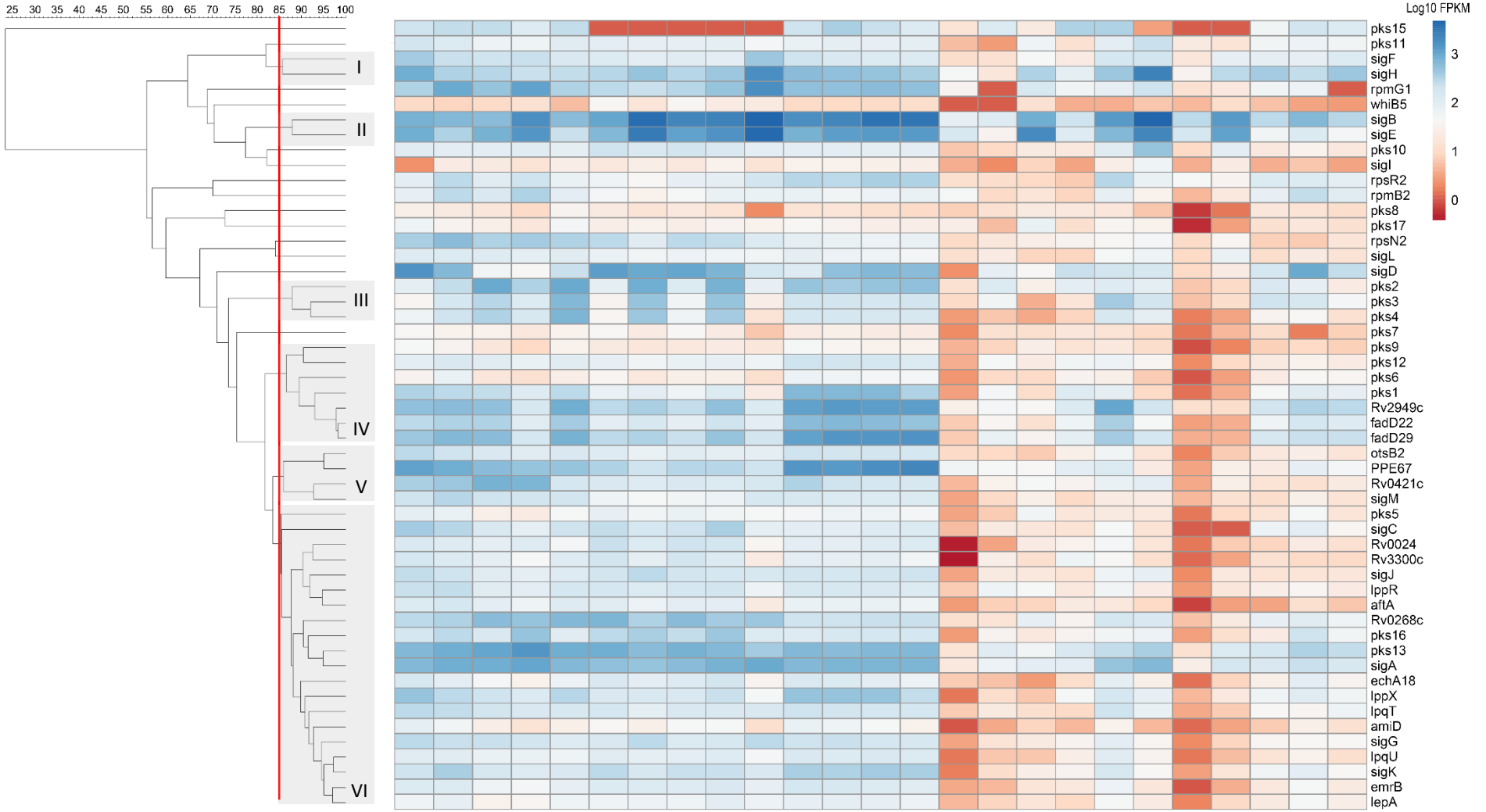
25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100
I
II
III
IV
V
VI
Log10 FPKM
3
2
1
0
pks15
pks11
sigF
sigH
rpmG1
whiB5
sigB
sigE
pks10
sigI
rpsR2
rpmB2
pks8
pks17
rpsN2
sigL
sigD
pks2
pks3
pks4
pks7
pks9
pks12
pks6
pks1
Rv2949c
fadD22
fadD29
otsB2
PPE67
Rv0421c
sigM
pks5
sigC
Rv0024
Rv3300c
sigJ
lppR
aftA
Rv0268c
pks16
pks13
sigA
echA18
lppX
lpqT
amiD
sigG
lpqU
sigK
emrB
lepA
HN878
H37Rv−HI
H37Rv−LI−1day
H37Rv−LI−1week
H37Rv−Glu
CDC−Gly−pH 7
CDC−Gly−pH 5.7
CDC−Pyr−pH 7
CDC−Pyr−pH 5.7
H37Rv−hypoxia
H37Rv−aera−1day
H37Rv−aera−2day
H37Rv−aera−3day
H37Rv−aera−4day
H37RV−FS
H37RV−FE
H37RV−DS
H37RV−DE
H37Rv−Log
H37Rv−Early
H37Rv−Medium
H37Rv−Late
Extract N0031
Extract N0153
Extract N0145

Figure 3.9 – Expression profiling of selected genes from *Mtb*, presented as log_{10} FPKM. Cut-off: 85% of similarity. Abbreviations: HN878 - *Mtb* HN878; H37Rv-HI - *Mtb* H37Rv - grown in high iron concentration; H37Rv-LI-1day - *Mtb* H37Rv grown in low iron concentration for 1 day; H37Rv-LI-1week- *Mtb* H37Rv grown in low iron concentration for 1 week; H37Rv-Glu - *Mtb* H37Rv grown in 0.4% glucose; CDC-Gly-pH 7 - *Mtb* CDC1551 grown in glycerol at pH 7; CDC-Gly-pH 5.7 - *Mtb* CDC1551 grown in glycerol at pH 5.7; CDC-Pyr-pH 7 - *Mtb* CDC1551 grown in pyruvate at pH 7; CDC-Pyr-pH 5.7 - *Mtb* CDC1551 grown in pyruvate at pH 5.7; H37Rv-hypoxia - *Mtb* H37Rv grown in hypoxia; H37Rv-aera-(1-4)day - *Mtb* H37Rv (1-4) day(s) after reaeration; H37RV-FS - *Mtb* H37Rv grown in long fatty acids at stationary phase; H37RV-FE - *Mtb* H37Rv grown in long fatty acids at exponential phase; H37RV-DS - *Mtb* H37Rv grown in dextrose at stationary phase; H37RV-DE - *Mtb* H37Rv grown in dextrose at exponential phase; H37Rv-Log - *Mtb* H37Rv in log phase; H37Rv-Early - *Mtb* H37Rv in early dormancy phase; H37Rv-Medium - *Mtb* H37Rv in medium dormancy phase; H37Rv-Late - *Mtb* H37Rv in late dormancy phase; Extract N0031 - *Mtb* clinical strain N0031; Extract N0153 - *Mtb* clinical strain N0153; and Extract N0145 - *Mtb* clinical strain N0145.

In this network, it becomes evident that *pks1* is highly correlated with *fadD22* (0.949), *Rv2949c* (0.910), *fadD29* (0.930) and also with *pks6* (0.920) and *pks12* (0.913). In agreement with clustering analysis, *pks4* is highly correlated with *pks3* (0.921) and *pks2* (0.910). Surprisingly, some FPKM values registered for *pks15* are nule, may be due to absence of mapped reads, thus it is not plotted in this network. A correlation between σ factors could also be confirmed. The *sigA* gene, known as the housekeeping regulator, is correlated with *sigG* (0.902), *sigJ* (0.902), *sigK* (0.902) and *sigM* (0.901); *sigG* is correlated with *sigJ* (0.938); and *sigK* (0.948) and *sigJ* with *sigK* (0.912). At the defined threshold, 30% of the correlations were found between members of bicluster module 0490, while for members of module 0211, correlations were only found between the genes of interest, *pks1* and *pks15*.

Both analyses provide indication that *pks1* expression shares high correlation with the expression profiles of *fadD22*, *Rv2949c* and *fadD29*, in agreement with reports from microarray data[72]. The expression pattern of *pks15* revealed by this analysis is impressively different from the one found in *pks1* in some of the stress conditions under examination, turning *pks15* into a single member cluster placed apart from the remaining clusters and, consequently, being absent from the correlation network. By contrast, microarray data available online shows that *pks15* is also highly correlated with *pks1* and *fadD22*. This fact may be due to the short half-life of most RNAs and subsequent mRNA degradation, which will greatly affect the output of RNA-seq. Our analyses also suggest that *pks4* is correlated with *pks3*, thus agreeing with previously published data[73] that defined that a *Mtb* H37Rv double mutant for *pks4* and *pks3* is not able to produce mycolipanoic, mycolipenic, and mycolipodienoic acids.

Also, the fact that *pks3* and *pks4* form a polyketide structure similar to the one verified in *pks15* and *pks1*, wherein *pks3* and *pks15* encode the ketoacyl synthase domain and *pks4* and *pks1* encode the remaining polyketide synthase domains, may indicate that *pks15* and *pks1* should be highly correlated, although this could not be confirmed by our data analysis. The *pks2* gene encodes a polyketide synthase that will synthetize methyl-branched fatty acids required for sulpholipid synthesis and belongs to one of the bicluster modules where are also included *pks4* and *pks3*[74].

Even though gene expression does not necessarily represent the activity of a specific σ factor, we also integrated this correlation network with their expression data to plot a representation of the putative regulation by σ factors (Figure 3.10). Six of the thirteen genes under analysis are highly correlated. Both *sigA*, the principal σ factor of mycobacteria, and *sigM*, responsible for long-term adaptation, are known to be regulated by *sigC*, whose role is to control virulence and immunopathology. The *sigA* factor is also known to regulate *sigG*, mostly induced during macrophage infection, which will regulate *sigJ*, known to be overexpressed in late stationary phase of dormant cultures. The *sigG* and *sigJ* will further regulate *sigL*, known to be involved in PDIM biosynthesis, that ahead regulate *sigK*, whose role remains partially unclear[44].

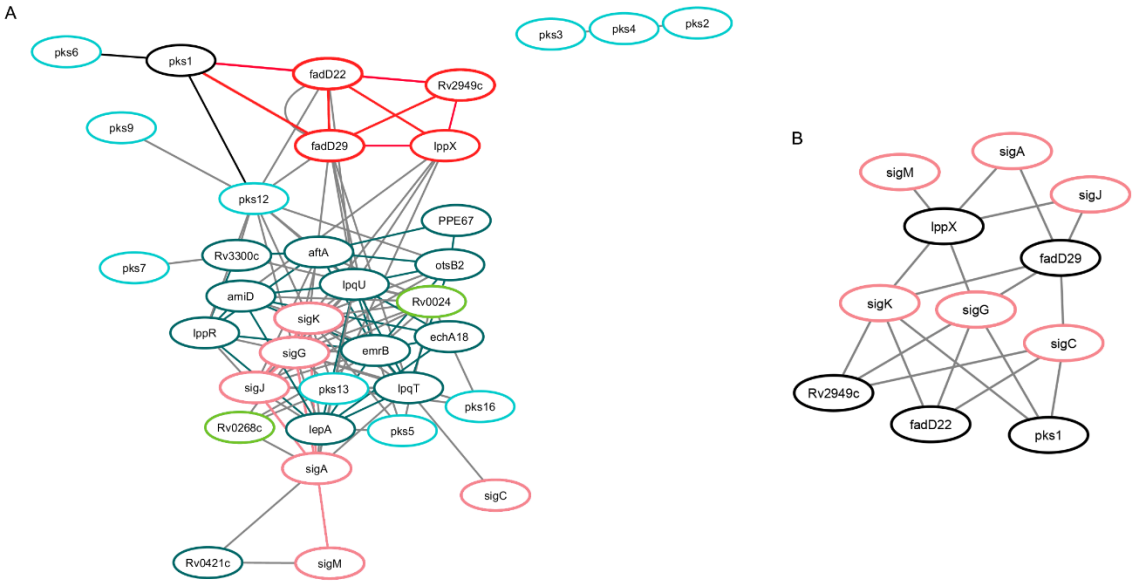

Figure 3.10 – Correlation network of expression data of selected genes for *Mtb*. A - Correlation threshold=0.9; in black: *pks1* and correlations to paralogs; in red: members of putative polycistronic structure and correlations between them and with *pks1*; in turquoise: genes encoding PKS; in pink: genes encoding σ factors and correlations between them; in dark blue: members of bicluster module 0490 and correlations between them; and in green: members of bicluster module 0211**.** B – Correlation threshold=0.8; In black: genes of interest; in pink: genes encoding σ factors.

None of the correlations with sigma factors plotted here involves directly our genes of interest, mainly due to the high threshold established for visualization on this correlation network. When we decreased the threshold basis, namely 0.8, we found that *pks1*, *Rv2949c*, *fadD22* and *fadD29* were all correlated with *sigC*, *sigG* and *sigK* (Figure 3.11). The *sigK* factor is predicted by *in silico* analysis to positively regulate *pks1* and *pks15*[75]. With this threshold, it becomes clearer that *sigA* and *sigJ* are correlated with *fadD29* and *lppX* and that *sigM* and *lppX* are also correlated with each other (Table 3.1). On contrary, amongst the analysis focused on sigma factors, *sigE* was the factor that presented the lowest correlations with the established genes of interest (*pks1*, 0.113; *fadD22*, 0.215; *Rv2949c*, 0.237; and *fadD29*, 0.267).

Table 3.1 – Pearson correlation coefficient between the selected genes of interest and genes encoding σ factors using a threshold of 0.8.

Gene 1	Gene 2	Pearson correlation coefficient
lppX	*sigA*	0.885
	sigG	0.886
	sigJ	0.851
	sigK	0.939
	sigM	0.817
pks1	*sigC*	0.811
	sigG	0.830
	sigK	0.850
fadD22	*sigC*	0.848
	sigG	0.877
	sigK	0.888
Rv2949c	*sigC*	0.824
	sigG	0.836
	sigK	0.873
fadD29	*sigA*	0.811
	sigC	0.864
	sigG	0.887
	sigJ	0.822
	sigK	0.909

For *Mb*, we also used 85% similarity as cut-off for cluster formation, but the pattern was a bit different; 15 clusters were generated, of which seven were single member clusters. Among the remaining eight clusters, we found cluster II including all members of the putative polycistronic structure, alongside with *Mb0429c* and *pks13*. We could observe cluster VIII composed by several PKS, namely *pks17*, *pks9*, *pks7*, *pks8*, *pks10* and *pks3*, alongside with the transcriptional regulator *whiB5* (Fig. 3.12). Those differences in cluster formation may be due to the smaller numbers of experiments used for analysis of *Mb* gene correlation, leading to less clusters but comprising a larger number of members.

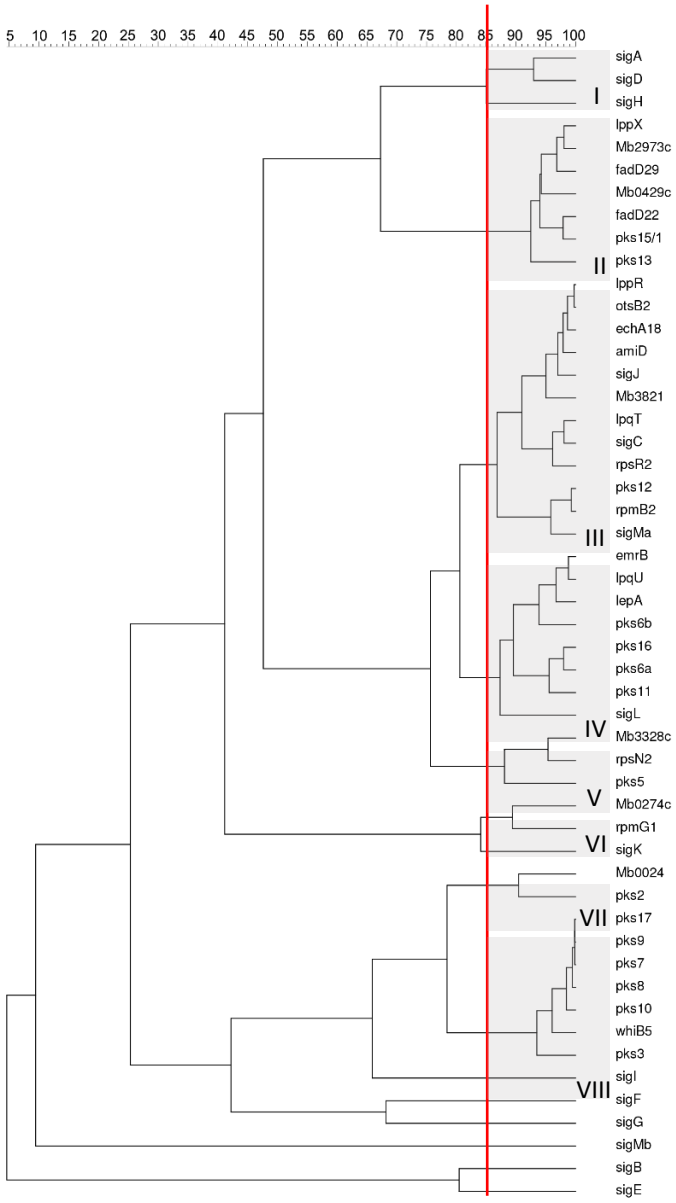

Figure 3.11 – Dendrogram from expression data of selected genes of *Mb* using as cut-off 85% similarity. Clustering analysis was performed based on six experimental conditions from two datasets.

As performed for *Mtb*, we constructed a correlation network with pairs of genes exhibiting correlation factors above 0.9 (Figure 3.12). On contrary to what we verified for *Mtb*, for *Mb* only *lppX* and *fadD22* present correlation values with *pks15/1* above the defined threshold. Also, it was verified a correlation above 0.9 for *lppX* with both *Mb2973c* and *fadD29*. Besides the correlations described above, *pks15/1* was also showed to be correlated with *sigM*. In agreement with the clustering analysis, the correlation pattern is very dissimilar from the one found for *Mtb*, may be due to a less robust set of data.

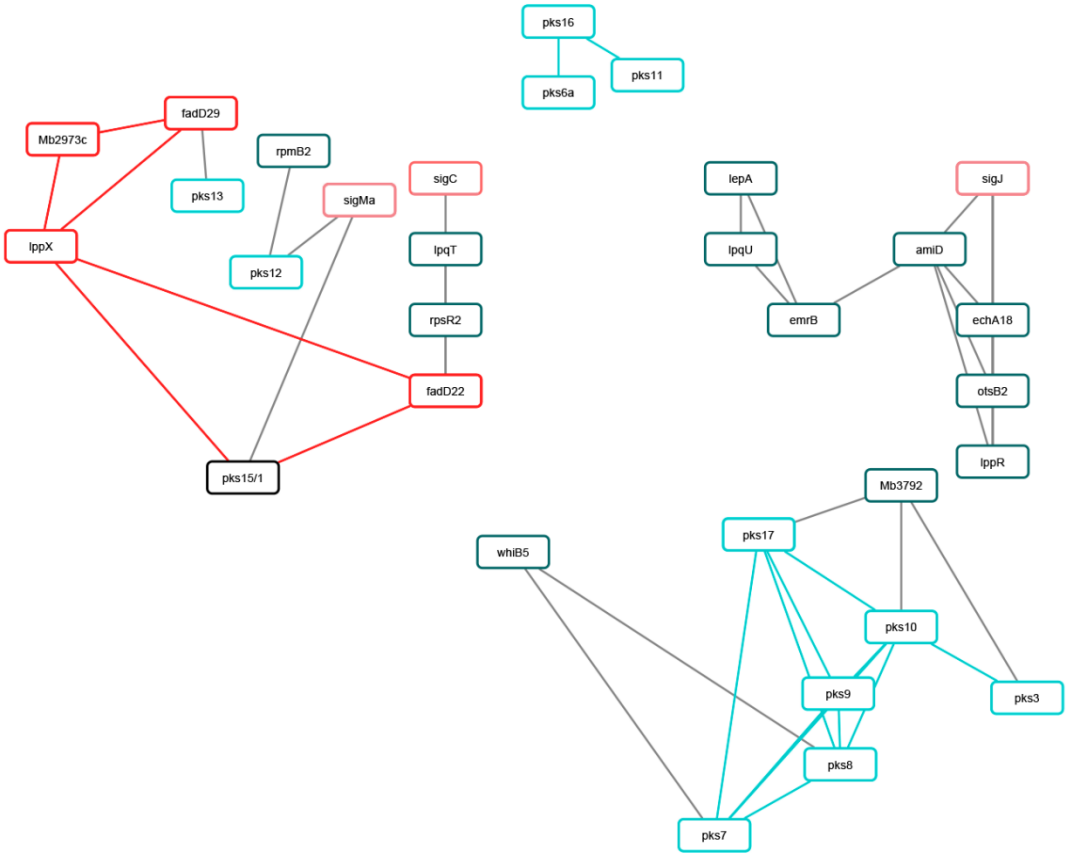

Figure 3.12 – Correlation network of expression data of selected genes for *Mb* using a correlation threshold of 0.9. In black: *pks15/1*; in red: members of putative polycistronic structure and correlations between them and with *pks15/1*; in turquoise: genes encoding PKS and correlations between them; in pink: genes encoding σ factors; and in dark blue: members of bicluster module 0490.

3.3.3 Differential expression of selected genes of interest

For differential expression inference, we analysed data using a method that takes read count data as input, uses negative binomial as reference distribution and normalizes data for analysis[61]. In this topic, we only analysed the set of genes putatively belonging to a polycistronic structure (*lppX*, *pks1*, *pks15*, *fadD22*, *Rv2949c* and *fadD29*). Also, only comparisons that represent stress conditions suitable to be found in *in vivo* growth were analysed in terms of differential expression.

Both for glycerol and pyruvate as carbon sources, *lppX*, *pks1*, *pks15*, *fadD22*, *Rv2949c* and *fadD29* are significantly down-regulated *in vitro* at pH 7, in contrast with the *in vivo* mimicking condition at pH 5.7. In the culture grown in pyruvate, $\log_2$ fold change values were found to be extremely significant and higher than in the sample grown in glycerol. When comparing carbon sources, there is no significant difference in expression of the selected genes of interest at pH 7.

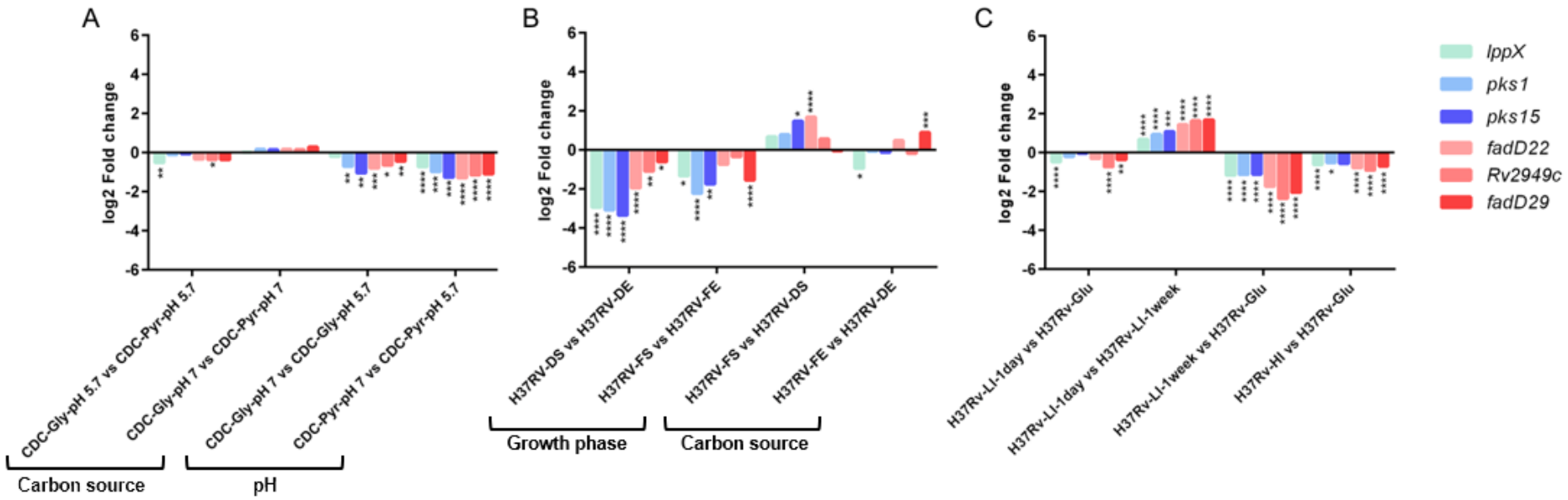

Figure 3.13 – Differential gene expression represented in log_2 fold change for *Mtb* strains. A – *Mtb* CDC1515 strain, pH and carbon source assay; B – *Mtb* H37Rv strain, growth phase and carbon source assay; C – *Mtb* H37Rv strain, iron concentration assay. Abbreviations: CDC-Gly-pH 7 - *Mtb* CDC1551 grown in glycerol at pH 7; CDC-Gly-pH 5.7 - *Mtb* CDC1551 grown in glycerol at pH 5.7; CDC-Pyr-pH 7 - *Mtb* CDC1551 grown in pyruvate at pH 7; CDC-Pyr-pH 5.7 - *Mtb* CDC1551 grown in pyruvate at pH 5.7; H37RV-FS - *Mtb* H37Rv grown in long fatty acids at stationary phase; H37RV-FE - *Mtb* H37Rv grown in long fatty acids at exponential phase; H37RV-DS - *Mtb* H37Rv grown in dextrose at stationary phase; H37RV-DE - *Mtb* H37Rv grown in dextrose at exponential phase; H37Rv-HI - *Mtb* H37Rv - grown in high iron concentration; H37Rv-LI-1day - *Mtb* H37Rv grown in low iron concentration for 1 day; H37Rv-LI-1week- *Mtb* H37Rv grown in low iron concentration for 1 week; H37Rv-Glu - *Mtb* H37Rv grown in 0.4% glucose. Statistical analysis was performed using Wald statistic test with p-value adjusted for multiple testing with the Benjamini-Hochberg procedure (α=0.05). * - significant (p-value=0.01 to 0.05), ** - very significant (p-value=0.001 to 0.01), *** - extremely significant (p-value=0.0001 to 0.001), **** - extremely significant (p-value< 0.0001).

However, at pH 5.7, a very significant difference in *lppX* expression level and a significant difference in *Rv2949c* were found, meaning that those genes are slightly downregulated in the sample grown in glycerol as sole carbon source (Figure 3.13 A). As mentioned above the conditions here explored allow the comparison between basal *in vitro* growth and *in vivo* growth inside phagosomes, wherein pH is lower. It is known that mycobacterial cell wall represents a major barrier to the entry of protons from its surrounds due to its complex structure. Also, it is known that most of the acid-sensitive mutants identified in *Mtb* present defects in genes involved in cell wall functions and several cell wall and lipid biosynthesis are known to be regulated by exposure to low pH[25]. As previously reported for *pks2*, *pks3* and *pks4*[26], we also confirm that *pks1*, *pks15* and the remaining members of the putative polycistronic structure, namely *lppX*, *fadD22*, *Rv2949c* and *fadD29*, are induced by exposure to acidic pH.

The comparisons between growth stages and carbon sources pointed out that, in both carbon sources, *pks1*, *pks15*, *fadD22*, *Rv2949c* and *fadD29*, are downregulated in stationary phase, when compared with exponential phase. In the cells grown in long fatty acids, only *pks1* and *fadD29* present extremely significant downregulation in stationary phase, being that *lppX* and *pks15* also present significant fold changes. By contrast, when bacteria were grown in dextrose, the complete set of genes under analysis was extremely significantly down-regulated in the stationary phase, except for *Rv2949c* and *fadD29* that present lower levels of significance. When comparing carbon sources for the same growth stages, *fadD29* appears to have an extremely significant up-regulation in long fatty acids at exponential phase, as well as *Rv2949c* which is extremely significantly up-regulated in long fatty acids in stationary phase (Figure 3.13 B). As expected, since we focused on genes that are part of the biosynthetic pathway of PGL, a significant downregulation is observed when the cultures enter the stationary phase, which may be explained by the fact that synthesis of the cell wall components is reduced at that time point. Comparing dextrose, the regular carbon source used *in vitro*, with growth in long fatty acids mimicking the triacylglycerols available at human cells[76], we could only identify the significant upregulation of *fadD29*, in exponential phase, and *pks15* and *fadD22*, in stationary phase, in the cells grown in long fatty acids.

In the iron concentration assays, it was possible to observe that, after 1 day of exposure to low iron concentration, only *lppX*, *Rv2949c* and *fadD29* were significantly differentially expressed, while after 1 week of exposure, the six genes were extremely significantly down-regulated when compared to the culture grown in 0.4% glucose. Those results are also supported by the direct comparison between the two cultures exposed to low iron concentration for different periods of time, that show extremely significant up-regulation in the culture exposed for 1 day. Similar to results from low iron concentration, exposure to high iron concentration also showed that *lppX*, *fadD22*, *Rv2949c* and *fadD29* are extremely significantly down-regulated and *pks1* is significantly down-regulated (Figure 3.13 C). We report that the selected set of genes is highly down-regulated under low iron concentrations, which could be related with the fact that iron takes part in several biological processes inside the cell, being required for cytochromes and other hemoproteins involved in oxygen metabolism. That means that iron deprivation can affect essential cellular processes, inducing a non-replicating state, thus reducing synthesis of cell wall components[24].

Comparing results obtained when cells were grown in hypoxia with the first 4 days after reaeration, it was possible to verify that *pks1*, *pks15*, *fadD22*, *Rv2949c* and *fadD29* were extremely significantly down-regulated in hypoxia, with some of the higher $\log_2$ fold change values verified among all assays. When comparing cells collected in the first, second and third days to those collected in the fourth day of reaeration, we found that both *lppX* and *pks1* have extremely significant $\log_2$ fold changes. Also, *fadD29* was found to be down-regulated with extremely significant differences from the first to the third and fourth days (Figure 3.14 A).

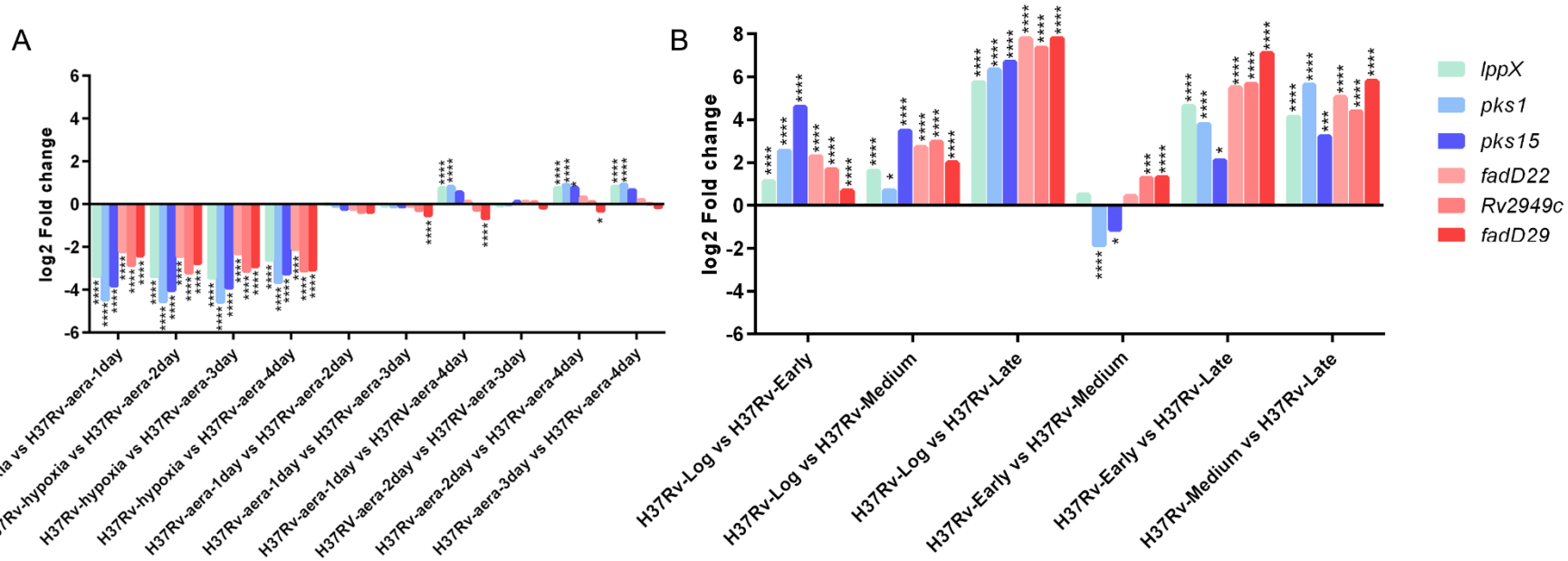

Figure 3.14 – Differential gene expression represented in log_2 fold change for *Mtb* strains. A – *Mtb* H37RV strain, hypoxia assay; and B – *Mtb* H37Rv strain, dormancy assay. Abbreviations: H37Rv-hypoxia - *Mtb* H37Rv grown in hypoxia; H37Rv-aera-(1-4)day - *Mtb* H37Rv (1-4) day(s) after reaeration; H37Rv-Log - *Mtb* H37Rv in log phase; H37Rv-Early - *Mtb* H37Rv in early dormancy phase; H37Rv-Medium - *Mtb* H37Rv in medium dormancy phase; H37Rv-Late - *Mtb* H37Rv in late dormancy phase. Statistical analysis was performed using Wald statistic test with p-value adjusted for multiple testing with the Benjamini-Hochberg procedure (α=0.05). * - significant (p-value=0.01 to 0.05), ** - very significant (p-value=0.001 to 0.01), *** - extremely significant (p-value=0.0001 to 0.001), **** - extremely significant (p-value< 0.0001).

Hypoxia induces many changes in mycobacteria. Both in microaerophilic and anaerobic cultures, *Mtb* is known to develop a thickened cell wall which may be important for adaptation to low-oxygen conditions[77]. However, the selected set of genes is showed to be extremely down-regulated in hypoxia, agreeing with previously published data[77], indicating that maybe the cell wall thickening does not occur in the outer layer.

Dormancy was induced in the experiments gathered by growing *Mtb* in K^+-deficient medium and, after 14–15 days of culture, adding rifampicin (5 μg/ml) to eliminate dividing bacteria. By comparing cells grown to three different states of dormancy (Table 2.1) with a culture grown to log phase in regular conditions, we got the higher fold changes verified in all assays. That comparison showed extremely significant down-regulation in dormancy conditions for all genes from the defined set. When comparing between states of dormancy, it was possible to find that *pks1* and *pks15* are down-regulated in early dormancy when comparing with medium dormancy. Also, for all genes, extremely significant fold changes were found between medium and late dormancy (Figure 3.14 B). While in dormancy, mycobacteria enter a state of low metabolic activity with alteration of gene regulation in order to accumulate triacylglycerols, loss of acid-fastness and a slower growth rate, which can thus explain why the selected genes of interest show a strong down-regulation under dormancy[77].

For *Mb*, we only analysed a starvation assay during three time points. It was possible to define that, when compared to exponential phase, all genes from the selected set are extremely significantly down-regulated along the three time-points. When comparing the expression in exponential phase with expression after resuscitation, only *lppX* was significantly down-regulated in exponential phase. Between time points, extremely significant fold changes were evident, except for the comparison between the tenth and the twentieth day (Figure 3.15).

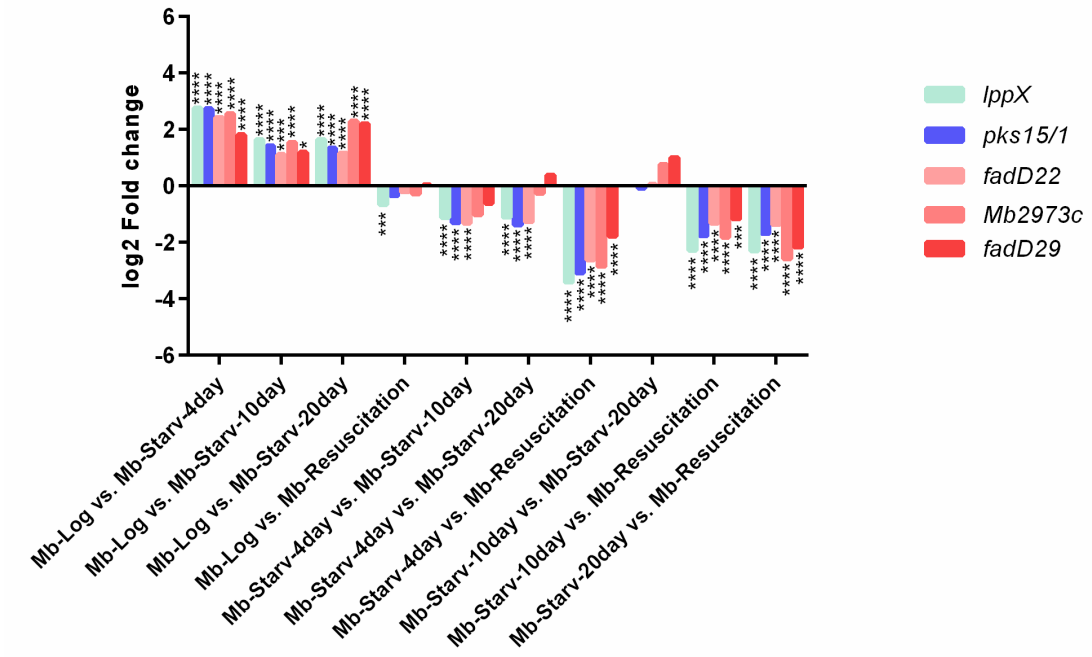

Figure 3.15 – Differential gene expression represented in $\log_2$ fold change for *Mb* strains. Starvation assay. Abbreviations: *Mb*-Log – *Mb* grown to log phase; *Mb*-Starv-(4, 10, 20) day – *Mb* grown during (4, 10, 20) days of starvation; and *Mb*-Resuscitation – *Mb* after resuscitation Statistical analysis was performed using Wald statistic test with p-value adjusted for multiple testing with the Benjamini-Hochberg procedure (α=0.05). * - significant (p-value=0.01 to 0.05), ** - very significant (p-value=0.001 to 0.01), *** - extremely significant (p-value=0.0001 to 0.001), **** - extremely significant (p-value< 0.0001).

4. Concluding Remarks and Future Perspectives

Mtb virulence is related to its aptitude to survive inside macrophages. During infection, macrophages engulf bacillus, generating a hostile intracellular environment for bacteria replication. *Mtb* interacts with macrophages in a critical and complex process, essential for the competition between pathogen and host. Recent models of persistency inside the host point to bacterial subpopulations in a latent state that maintain their ability to reactivate upon host's immunosuppression. One of the processes that is speculated to occur during infection is the development of a cell community structure and organization pattern similar to microbial biofilms. One of the most relevant elements of biofilm formation is the bacterial cell wall, which works as an interface with biotic or abiotic surfaces. The study of the regulation of glycolipids biosynthesis such as PGL is of special interest once PGL might be relevant for mycobacterial ability to form biofilms. PGL production involves several PKS, namely *pks1* and *pks15*, which have been shown to have a critical role in this pathway, since the presence of a mutation that disrupts *pks15/1* CDS was associated with lack of PGL production[41]. Also, it is known that the reference strain for pathogenic mycobacteria, *Mtb* H37Rv, presents such mutation. Our focus in this work was to unveil the transcriptional structure and function of *pks1* and *pks15* by cloning their putative regulatory regions in transcriptional fusion vectors and by attempting the construction of a *pks1* transposon-free knock-out mutant, respectively. Since previous studies in our research group have focused on relating biofilm formation under several growth conditions with *pks1* and *pks15* expression profile inferred by qPCR, in the present thesis we focused on inferring the regulatory pattern of both genes, using a genome-wide approach by analysis of transcriptome (RNA-seq) data.

The analysis of expression data gathered from publicly available sources suggested that the target genes selected for this work, *pks1* and *pks15*, may be transcribed as a polycistronic unit composed by three to six genes located both upstream of *pks15* and downstream of *pks1*. All these gene products except FadD29*'* take part of the biosynthetic pathway of phenolphtiocerol moiety of PGL. Also, *pks1* and *pks15* both seem to be positively regulated by *sigK* and negatively regulated by *sigE*, based on our algorithm prediction.

Cloning of putative regulatory regions was performed in pJET1.2/blunt for *pks1*, *pks15* and *fadD22*, the minimal structure proposed for this operon, and was successful for all three targets but, unfortunately, all of our attempts to subclone in the mycobacterial replicating vector, pSM128, were unsuccessful. This cloning step remains one of the key elements to clarify the transcriptional regulation of our genes of interest, once it would enable to determine the exact physical location of the operon promoter and to understand which environmental conditions trigger or inhibit this operon expression.

Another experimental strategy that did not work favourably was the attempt to generate a knock-out mutant of *pks1*, neither by phage-mediated mutagenesis or mycobacterial recombineering. By recombineering, we successfully made a recombineering strain with pJV53, although our attempts to electroporate the designed AES were unsuccessful. By phage-mediated mutagenesis, we were able to generate hypothetic transductants but, due to the high dimension of the desired transductant substrate (over 55 kb), we were not able to successfully verify if the lambda-derived phasmid vector was indeed ligated to the AES-carrier cosmid, using the available commercial kits.

By clustering the expression data from RNA-seq datasets for *Mtb*, in a robust set of 25 growth conditions, it was possible to relate *pks1* with *fadD22*, *Rv2949c*, *fadD29*, *pks6*, *pks12* and *pks9*, gathering evidence that the first three genes are part of a polycistronic structure. In another cluster, we also found *pks4* and *pks3* with a similarity percentage above 85%.

Those two PKS form a structure similar to *pks1* and *pks15*, where one of the genes encodes ketosynthase domain and the other encodes for the remaining domains. With a closer analysis, focused on the values of correlation coefficient, we were once again able to confirm that all genes thought to belong to the putative polycistronic structure present similar expression profiles. Correlations between those genes were shown to be above 0.9, except for *pks15*. Also, we here found that *pks1* only presented correlation coefficient values above 0.9 with *pks6* and *pks12*. Away from these levels of correlation, as referred, was *pks15*, even though it is mostly due to the presence of several nule FPKM values and not to the dissimilarity of the expression profile. In this integrating analysis, it was also possible to relate genes encoding σ factors with the selected genes of interest. Even though we have to carefully appreciate those results, since σ gene expression may not reflect factor activity, we found correlation coefficients above 0.8 between *sigC*, *sigG* and *sigK* and the members of the putative polycistronic structure. While for *Mtb*, we managed to gather a robust set of data, for *Mb* it was only possible to collect data from six growth conditions of which three represent regular *in vitro* growth and three represent growth under starvation at three time-points. This smaller data set led to lower number of clusters but with higher number of members. Although this data set bias may influence interpretation, it was possible to verify that all members of the putative polycistronic structure belong to the same cluster alongside with *Mb0429c* and *pks13*.

As referred, mycobacteria are subjected to several stress conditions while inside macrophages. By analysing the differential expression of *lppX*, *pks1*, *pks15*, *fadD22*, *Rv2949c* and *fadD29*, it was possible to define in which conditions are those genes positively or negatively regulated. We analysed expression levels under growth in six different conditions, namely pH, carbon source, growth phase, exposure to limiting or excessive iron concentration, hypoxia and dormancy, confirming that the selected genes of interest are up-regulated at acidic pH and down-regulated at stationary phase, under hypoxia and dormancy, and at both low and high iron concentrations (Fig. 4.1). The combination of both sets of data, clustering of genes by expression data and differential expression analysis, suggests *lppX* is isolated from the remaining set of genes. Also, in one of the conditions, *lppX*, *Rv2949c* and *fadD29* expression seemed to diverge from that of *pks1*, *pks15* and *fadD22* (Fig. 4.1). Using differential expression analysis, we were also able to confirm that *sigK* shares the expression profile with the selected genes of interest in 80% of the conditions with significant fold-changes; the same percentage is also verified for *sigD*. On the contrary, *sigE* presents approximately 78% of expression profile dissimilarity with the selected panel of genes in the conditions under analysis, with significant fold-changes, as well as *sigB*.

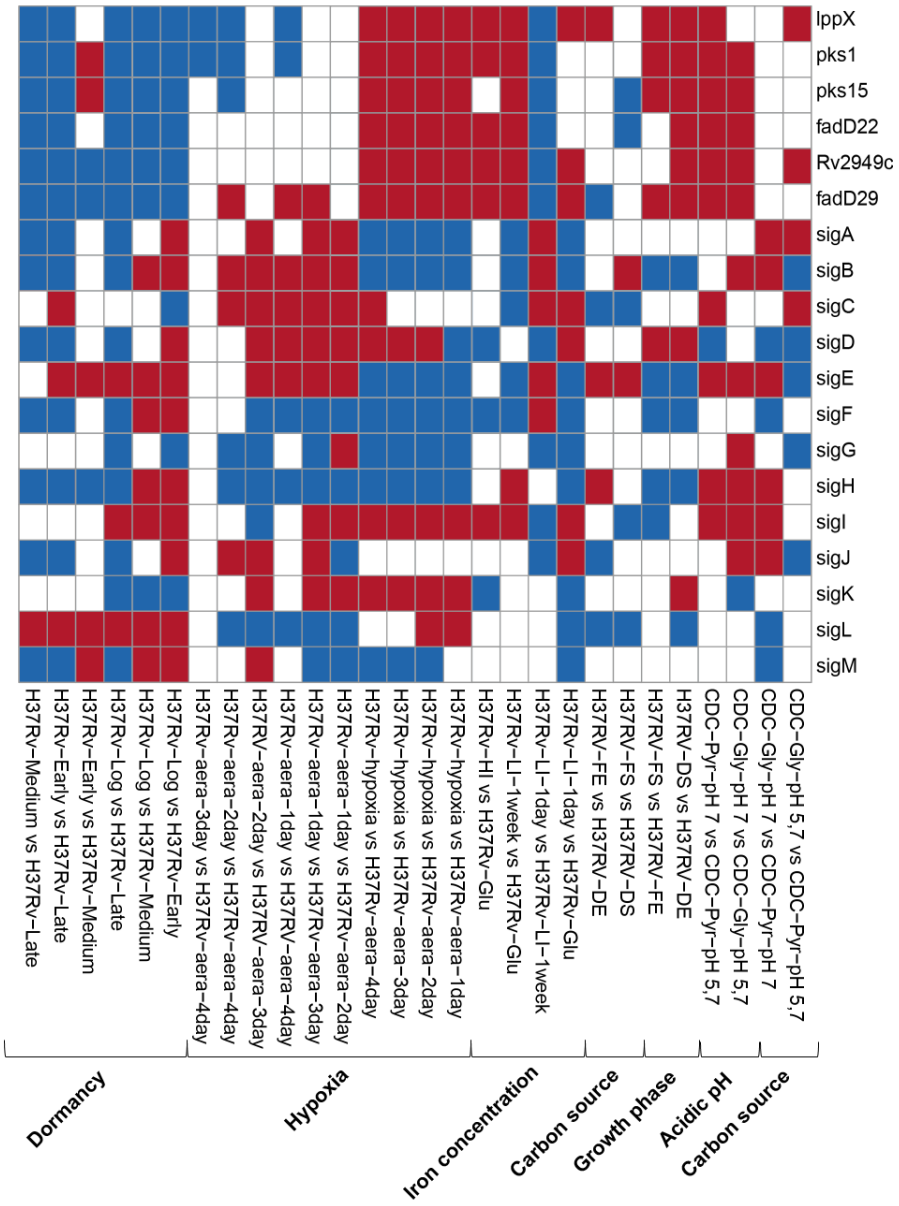

Figure 4.1 – Regulation pattern of selected genes of interest and genes encoding σ factors. In blue: upregulation; in red: downregulation; and in white: non-significant fold-changes.

Gathering the previously published data and the transcriptome data we were able to collect and analyse, we propose a regulatory model for *pks1* and *pks15* (Fig. 4.2). In this model, we used a conservative approach in which we only selected genes sharing total expression profiles and functional similarity in terms of the polycistronic structure model. For this model, we selected a set of three genes, *pks1*, *pks15* and *fadD22*, that fulfil the statements above (Fig. 4.2). Based on differential expression analysis, we also selected a set of four σ factors (σ^D and σ^K, and σ^B and σ^E) that putatively regulate the expression of *pks1*, *pks15* and *fadD22* (Fig. 4.2). Both σ^K and σ^E were previously predicted to regulate the genes belonging to the polycistronic structure under hypothesis[50], being that the corresponding expression data analysed in this work give support to that prediction. However, it was possible to find other two genes encoding σ factors with similar percentages of expression profile, namely σ^D and σ^B, that were also included in our model proposal. The genes encoding factors σ^D and σ^K were shown to be down-regulated under hypoxia and dormancy, as well as in stationary phase. On the contrary, the genes encoding σ^B and σ^E factors are up-regulated in the same conditions, being that *sigB* was previously shown to be up-

regulated under hypoxia[75]. While σ^D and σ^K, that appear to positively regulate the selected genes of interest, belong to the lower level of σ factors regulation, the σ^B and σ^E factors, that putatively regulate in a negative way the selected genes of interest, belong to upper levels of regulation.

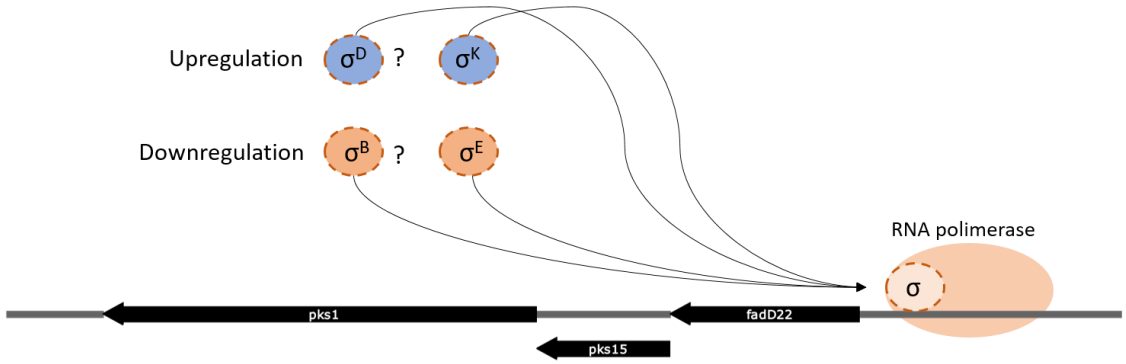

Figure 4.2 – Schematic representation of the proposed polycistronic structure model. *pks1*, *pks15* and *fadD22* are represented with putative regulation from σ^D and σ^K (positive), and σ^B and σ^E (negative).

As future perspectives, it would be extremely relevant the completion of one of the knock-out mutant protocols, so we could test *in vitro* the function of *pks1*, initially by comparing the original phenotype with the phenotype of the mutant strain in stress conditions, later by *in vitro* assays of infection in macrophages to evaluate intracellular survival of the mutant strain and, finally, by *in vivo* assays using an animal infection model to get insights into the virulence of the mutant strain and thus compare with the parental'. Also, it would be relevant to improve the construction of the transcriptional fusion of the upstream regions of the selected genes of interest with *lacZ* reporter gene by a high throughput method, once we were working with a low probability of positive ligation. This cloning represents a very important step towards unveiling the exact promoter location and activity. Also, it would be interesting to analyse *pks1* and *pks15* gene expression in mutant strains of their putative regulators under several growth conditions. All that regulatory data would be extremely relevant to refine knowledge on the regulation of *pks1* and *pks15*, previously shown to assume a crucial role in PGL production[40] and thus at the interface with the host.

5. References

1. **The Prokaryotes**, vol. 3, 3rd edn: Springer; 2006.
2. Madigan, MT, Martinko, JM, *et al.*: **Brock Biology of Microrganisms**, 14th edn: Pearson; 2015.
3. Levy-Frebault, VV, Portaels, F: **Proposed minimal standards for the genus *Mycobacterium* and for description of new slowly growing *Mycobacterium* species**. *International journal of systematic bacteriology* 1992, **42**(2):315-323.
4. Forrellad, MA, Klepp, LI, *et al.*: **Virulence factors of the *Mycobacterium tuberculosis* complex**. *Virulence* 2013, **4**(1):3-66.
5. Kiers, A, Klarenbeek, A, *et al.*: **Transmission of *Mycobacterium pinnipedii* to humans in a zoo with marine mammals**. *The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease* 2008, **12**(12):1469-1473.
6. Smith, NH, Crawshaw, T, *et al.*: ***Mycobacterium microti*: More diverse than previously thought**. *J Clin Microbiol* 2009, **47**(8):2551-2559.
7. Smith, NH, Hewinson, RG, *et al.*: **Myths and misconceptions: the origin and evolution of *Mycobacterium tuberculosis***. *Nat Rev Micro* 2009, **7**(7):537-544.
8. Brosch, R, Gordon, SV, *et al.*: **A new evolutionary scenario for the *Mycobacterium tuberculosis* complex**. *Proceedings of the National Academy of Sciences* 2002, **99**(6):3684-3689.
9. Bottai, D, Stinear, TP, *et al.*: **Mycobacterial Pathogenomics and Evolution**. *Microbiology spectrum* 2014, **2**(1):Mgm2-0025-2013.
10. Organization, WH: **Global Tuberculosis Report 2016**. In.; 2016.
11. ECDC: **Annual Epidemiological Report** [http://ecdc.europa.eu/en/healthtopics/Tuberculosis/Pages/Annual-epidemiological-report-2016.aspx]; 2014; Visited at: June 16, 2016.
12. Nunn, P, Williams, B, *et al.*: **Tuberculosis control in the era of HIV**. *Nat Rev Immunol* 2005, **5**(10):819-826.
13. Michel, AL, Muller, B, *et al.*: ***Mycobacterium bovis* at the animal-human interface: a problem, or not?** *Veterinary microbiology* 2010, **140**(3-4):371-381.
14. Health, CfFSaP: **Bovine Tuberculosis**. In.; 2009.
15. Pieters, J: **Entry and survival of pathogenic mycobacteria in macrophages**. *Microbes and Infection* 2001, **3**(3):249-255.
16. Sundaramurthy, V, Pieters, J: **Interactions of pathogenic mycobacteria with host macrophages**. *Microbes and Infection* 2007, **9**(14):1671-1679.
17. Korbel, DS, Schneider, BE, *et al.*: **Innate immunity in tuberculosis: myths and truth**. *Microbes Infect* 2008, **10**(9):995-1004.
18. Sia, JK, Georgieva, M, *et al.*: **Innate Immune Defenses in Human Tuberculosis: An Overview of the Interactions between *Mycobacterium tuberculosis* and Innate Immune Cells**. *Journal of immunology research* 2015, **2015**:747543.
19. Fogel, N: **Tuberculosis: a disease without boundaries**. *Tuberculosis (Edinburgh, Scotland)* 2015, **95**(5):527-531.
20. Huynh, KK, Joshi, SA, *et al.*: **A delicate dance: host response to mycobacteria**. *Current opinion in immunology* 2011, **23**(4):464-472.
21. Leistikow, RL, Morton, RA, *et al.*: **The *Mycobacterium tuberculosis* DosR regulon assists in metabolic homeostasis and enables rapid recovery from nonrespiring dormancy**. *J Bacteriol* 2010, **192**(6):1662-1670.
22. Sivaramakrishnan, S, Ortiz de Montellano, PR: **The DosS-DosT/DosR Mycobacterial Sensor System**. *Biosensors* 2013, **3**(3):259-282.
23. Neyrolles, O, Wolschendorf, F, *et al.*: **Mycobacteria, metals, and the macrophage**. *Immunological reviews* 2015, **264**(1):249-263.
24. Sritharan, M: **Iron Homeostasis in *Mycobacterium tuberculosis*: Mechanistic Insights into Siderophore-Mediated Iron Uptake**. *Journal of Bacteriology* 2016, **198**(18):2399-2409.

25. Vandal, OH, Nathan, CF, *et al.*: **Acid Resistance in *Mycobacterium tuberculosis***. *Journal of Bacteriology* 2009, **191**(15):4714-4721.
26. Rohde, KH, Abramovitch, RB, *et al.*: ***Mycobacterium tuberculosis* Invasion of Macrophages: Linking Bacterial Gene Expression to Environmental Cues**. *Cell Host & Microbe*, **2**(5):352-364.
27. Keating, LA, Wheeler, PR, *et al.*: **The pyruvate requirement of some members of the *Mycobacterium tuberculosis* complex is due to an inactive pyruvate kinase: implications for in vivo growth**. *Molecular microbiology* 2005, **56**(1):163-174.
28. Marrakchi, H, Lanéelle, M-A, *et al.*: **Mycolic Acids: Structures, Biosynthesis, and Beyond**. *Chemistry & Biology* 2014, **21**(1):67-85.
29. Angala, SK, Belardinelli, JM, *et al.*: **The cell envelope glycoconjugates of *Mycobacterium tuberculosis***. *Critical reviews in biochemistry and molecular biology* 2014, **49**(5):361-399.
30. Bailo, R, Bhatt, A, *et al.*: **Lipid transport in *Mycobacterium tuberculosis* and its implications in virulence and drug development**. *Biochemical pharmacology* 2015, **96**(3):159-167.
31. Chiaradia, L, Lefebvre, C, *et al.*: **Dissecting the mycobacterial cell envelope and defining the composition of the native mycomembrane**. *Scientific Reports* 2017, **7**(1):12807.
32. Onwueme, KC, Vos, CJ, *et al.*: **The dimycocerosate ester polyketide virulence factors of mycobacteria**. *Progress in lipid research* 2005, **44**(5):259-302.
33. Ferreras, JA, Stirrett, KL, *et al.*: **Mycobacterial PGL virulence factor biosynthesis: mechanism and small-molecule inhibition of polyketide chain initiation**. *Chemistry & biology* 2008, **15**(1):51-61.
34. Astarie-Dequeker, C, Le Guyader, L, *et al.*: **Phthiocerol Dimycocerosates of *M. tuberculosis* Participate in Macrophage Invasion by Inducing Changes in the Organization of Plasma Membrane Lipids**. *PLoS pathogens* 2009, **5**(2):e1000289.
35. Rousseau, C, Winter, N, *et al.*: **Production of phthiocerol dimycocerosates protects *Mycobacterium tuberculosis* from the cidal activity of reactive nitrogen intermediates produced by macrophages and modulates the early immune response to infection**. *Cellular microbiology* 2004, **6**(3):277-287.
36. Pang, JM, Layre, E, *et al.*: **The polyketide Pks1 contributes to biofilm formation in *Mycobacterium tuberculosis***. *J Bacteriol* 2012, **194**(3):715-721.
37. Kieser, KJ, Rubin, EJ: **How sisters grow apart: mycobacterial growth and division**. *Nature reviews Microbiology* 2014, **12**(8):550-562.
38. Quadri, LE: **Biosynthesis of mycobacterial lipids by polyketide synthases and beyond**. *Crit Rev Biochem Mol Biol* 2014, **49**(3):179-211.
39. Ridley, CP, Lee, HY, *et al.*: **Evolution of polyketide synthases in bacteria**. *Proceedings of the National Academy of Sciences of the United States of America* 2008, **105**(12):4595-4600.
40. He, W, Soll, CE, *et al.*: **Cooperation between a Coenzyme A-Independent Stand-Alone Initiation Module and an Iterative Type I Polyketide Synthase during Synthesis of Mycobacterial Phenolic Glycolipids**. *Journal of the American Chemical Society* 2009, **131**(46):16744-16750.
41. Constant, P, Perez, E, *et al.*: **Role of the pks15/1 gene in the biosynthesis of phenolglycolipids in the *Mycobacterium tuberculosis* complex. Evidence that all strains synthesize glycosylated p-hydroxybenzoic methyl esters and that strains devoid of phenolglycolipids harbor a frameshift mutation in the pks15/1 gene**. *The Journal of biological chemistry* 2002, **277**(41):38148-38158.
42. Daffe, M, Crick, DC, *et al.*: **Genetics of Capsular Polysaccharides and Cell Envelope (Glyco)lipids**. *Microbiology spectrum* 2014, **2**(4):Mgm2-0021-2013.
43. Flentie, K, Garner, AL, *et al.*: **The *Mycobacterium tuberculosis* transcription machinery: ready to respond to host attacks**. *Journal of Bacteriology* 2016.
44. Sachdeva, P, Misra, R, *et al.*: **The sigma factors of *Mycobacterium tuberculosis*: regulation of the regulators**. *FEBS Journal* 2010, **277**(3):605-626.

45. Sambrook, J, Russell, DW: **Molecular Cloning: A Laboratory Manual**, vol. 1: CSHL Press; 2001.
46. Lew, JM, Kapopoulou, A, *et al.*: **TubercuList--10 years after**. *Tuberculosis (Edinburgh, Scotland)* 2011, **91**(1):1-7.
47. **Database Resources of the National Center for Biotechnology Information**. *Nucleic acids research* 2017, **45**(D1):D12-D17.
48. Reddy, TB, Riley, R, *et al.*: **TB database: an integrated platform for tuberculosis research**. *Nucleic acids research* 2009, **37**(Database issue):D499-508.
49. Galagan, JE, Sisk, P, *et al.*: **TB database 2010: overview and update**. *Tuberculosis (Edinburgh, Scotland)* 2010, **90**(4):225-235.
50. Turkarslan, S, Peterson, EJ, *et al.*: **A comprehensive map of genome-wide gene regulation in *Mycobacterium tuberculosis***. *Scientific data* 2015, **2**:150010.
51. Hyatt, D, Chen, G-L, *et al.*: **Prodigal: prokaryotic gene recognition and translation initiation site identification**. *BMC bioinformatics* 2010, **11**(1):119.
52. Oberto, J: **SyntTax: a web server linking synteny to prokaryotic taxonomy**. *BMC bioinformatics* 2013, **14**:4.
53. Leinonen, R, Sugawara, H, *et al.*: **The Sequence Read Archive**. *Nucleic acids research* 2011, **39**(suppl_1):D19-D21.
54. Trapnell, C, Roberts, A, *et al.*: **Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks**. *Nat Protocols* 2012, **7**(3):562-578.
55. Kim, D, Pertea, G, *et al.*: **TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions**. *Genome Biology* 2013, **14**(4):R36.
56. Trapnell, C, Williams, BA, *et al.*: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation**. *Nat Biotech* 2010, **28**(5):511-515.
57. Metsalu, T, Vilo, J: **ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap**. *Nucleic acids research* 2015, **43**(W1):W566-570.
58. Smoot, ME, Ono, K, *et al.*: **Cytoscape 2.8: new features for data integration and network visualization**. *Bioinformatics* 2011, **27**(3):431-432.
59. Saito, R, Smoot, ME, *et al.*: **A travel guide to Cytoscape plugins**. *Nat Methods* 2012, **9**(11):1069-1076.
60. Anders, S, Pyl, PT, *et al.*: **HTSeq--a Python framework to work with high-throughput sequencing data**. *Bioinformatics* 2015, **31**(2):166-169.
61. Love, MI, Huber, W, *et al.*: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2**. *Genome Biol* 2014, **15**(12):550.
62. Afgan, E, Baker, D, *et al.*: **The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update**. *Nucleic acids research* 2016, **44**(W1):W3-W10.
63. Dussurget, O, Timm, J, *et al.*: **Transcriptional Control of the Iron-ResponsivefxbA Gene by the Mycobacterial Regulator IdeR**. *Journal of Bacteriology* 1999, **181**(11):3402-3408.
64. Casadaban, MJ, Cohen, SN: **Analysis of gene control signals by DNA fusion and cloning in Escherichia coli**. *Journal of molecular biology* 1980, **138**(2):179-207.
65. van Kessel, JC, Hatfull, GF: **Mycobacterial Recombineering**. In: *Chromosomal Mutagenesis.* Edited by Davis, GD, Kayser, KJ. Totowa, NJ: Humana Press; 2008: 203-215.
66. Kalpana, GV, Bloom, BR, *et al.*: **Insertional mutagenesis and illegitimate recombination in mycobacteria**. *Proceedings of the National Academy of Sciences* 1991, **88**(12):5433-5437.
67. Tufariello, JM, Malek, AA, *et al.*: **Enhanced Specialized Transduction Using Recombineering in *Mycobacterium tuberculosis***. *mBio* 2014, **5**(3).
68. van Kessel, JC, Hatfull, GF: **Mycobacterial recombineering**. *Methods in molecular biology (Clifton, NJ)* 2008, **435**:203-215.

69. Reiss, DJ, Baliga, NS, *et al.*: **Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks**. *BMC bioinformatics* 2006, **7**:280.
70. Mawuenyega, KG, Forst, CV, *et al.*: ***Mycobacterium tuberculosis* functional network analysis by global subcellular protein profiling**. *Molecular biology of the cell* 2005, **16**(1):396-404.
71. Griffin, JE, Gawronski, JD, *et al.*: **High-resolution phenotypic profiling defines genes essential for mycobacterial growth and cholesterol catabolism**. *PLoS pathogens* 2011, **7**(9):e1002251.
72. Dehal, PS, Joachimiak, MP, *et al.*: **MicrobesOnline: an integrated portal for comparative and functional genomics**. *Nucleic acids research* 2010, **38**(Database issue):D396-D400.
73. Dubey, VS, Sirakova, TD, *et al.*: **Disruption of msl3 abolishes the synthesis of mycolipanoic and mycolipenic acids required for polyacyltrehalose synthesis in Mycobacterium tuberculosis H37Rv and causes cell aggregation**. *Molecular microbiology* 2002, **45**(5):1451-1459.
74. Sirakova, TD, Thirumala, AK, *et al.*: **The *Mycobacterium tuberculosis* pks2 gene encodes the synthase for the hepta- and octamethyl-branched fatty acids required for sulfolipid synthesis**. *The Journal of biological chemistry* 2001, **276**(20):16833-16839.
75. Bonneau, R, Reiss, DJ, *et al.*: **The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo**. *Genome Biology* 2006, **7**(5):R36-R36.
76. Rodríguez, JG, Hernández, AC, *et al.*: **Global Adaptation to a Lipid Environment Triggers the Dormancy-Related Phenotype of *Mycobacterium tuberculosis***. *mBio* 2014, **5**(3).
77. Alnimr, AM: **Dormancy models for *Mycobacterium tuberculosis*: A minireview**. *Brazilian Journal of Microbiology* 2015, **46**(3):641-647.
78. Snapper, SB, Melton, RE, *et al.*: **Isolation and characterization of efficient plasmid transformation mutants of *Mycobacterium smegmatis***. *Molecular microbiology* 1990, **4**(11):1911-1919.
79. Bardarov, S, Bardarov Jr, S, Jr., *et al.*: **Specialized transduction: an efficient method for generating marked and unmarked targeted gene disruptions in *Mycobacterium tuberculosis, M. bovis* BCG and *M. smegmatis***. *Microbiology (Reading, England)* 2002, **148**(Pt 10):3007-3017.
80. Prata, A: **Estudo dos determinantes ambientais e genéticos que promovem o crescimento de bactérias do complexo *Mycobacterium tuberculosis* em biofilme: o papel dos genes *pks1* e *pks15***. *MSc Thesis.* Faculdade de Ciências, Universidade de Lisboa; 2016.

Supplementary Table 6.1 – List of bacterial strains used in the current work, with relevant phenotype and reference/origin.

Strain	Relevant Phenotype	Reference or origin
Mycobacterium smegmatis mc^2 155	Ept+ , KanS	[78]
E. coli αDH5	endA1 hsdR17(rk- , mk+) supE44 thi -1 recA1 gyrA96 relA1 lac [F´ proA+B + lacIq ZΔM15:Tn10(TcR)]	NZYTech
E. coli HB101	F - (gpt-proA)62, leuB6, glnV44, ara-14, galK2, lacY1, (mcrC-mrr), rpsL20(Strr), xyl-5, mtl-1, recA13, thi-1	INIAV
E. coli MC1061	araD139 Δ(araA-leu)7697 Δ(lac)X74 galK16 galE15(GalS) lambda- e14- mcrA0 relA1 rpsL150(strR) spoT1 mcrB1 hsdR2	FCUL

Mutations description:

Ept+ - efficient plasmid transformation phenotype;

KanS – kanamycin resistant;

endA1 - allows cleaner preparations of DNA and better results in downstream applications due to the elimination of non-specific digestion by Endonuclease I;

hsdR17(rk- , mk+) - eliminates the restriction endonuclease of the EcoKI restriction-modification system, so DNA lacking the EcoKI methylation will not be degraded;

supE44 - suppression of UAG stop codons by insertion of glutamine; required for some phage growth;

thi -1 - requires thiamine;

recA1 - reduces occurrence of unwanted recombination in cloned DNA; cells UV sensitive, deficient in DNA repair;

gyrA96 - mutation in DNA gyrase; conveys nalidixic acid resistance;

relA1 - permits RNA synthesis in absence of protein synthesis;

lac - deletion of the entire lac operon;

[F´ proA+B + lacIq ZΔM15:Tn10(TcR)] - contains an F' episomal plasmid with the stated features: the genes proA and proB encoding the first two enzymes of the proline biosynthetic sequence in *Escherichia coli*; overproduction of the lac repressor protein; strain carries the lacZ deletion mutant which contains the ω-peptide: a mutant β-galactosidase derived from the M15 strain of E. coli that has its N-terminal residues 11—41 deleted and is unable to form a tetramer so it is inactive; and transposon normally carrying tetracycline resistance;

F- - strain does not carry the F plasmid;

(gpt-proA)62 – strain can not synthesize proline;

leuB6 - requires leucine;

glnV44 - suppression of UAG stop codons by insertion of glutamine; required for some phage growth;

ara-14 - cannot metabolize arabinose;

galK2 - mutants cannot metabolize galactose and are resistant to 2-deoxygalactose;

lacY1 - deficient in lactose transport; deletion of lactose permease;

(mcrC-mrr) - deletes six genes in the restriction system that requires methyl mcrBC cytosine is abolished;

rpsL20(StrR) - mutation in ribosomal protein S12 conveying streptomycin resistance;

xyl-5 - blocked xylose metabolism;

mtl-1 - strain has a mutation in a gene evolved in the mannitol metabolism pathway (the mannitol utilization is blocked);

recA13 - as for recA1, but inserts less stable;

thi-1 - requires thiamine;

araD139 - mutation in L-ribulose-phosphate 4-epimerase blocks arabinose metabolism;

Δ(araA-leu)7697 - deletion of the first 295 codons of hepA to the first 82 codons of fruR;

Δ(lac)X74 - deletion of the entire lac operon as well as some flanking DNA;

galK16 - is an IS2 insertion ~170bp downstream of the galK start codon;

galE15- is a point mutation resulting in a Ser123 -> Phe conversion near the enzyme's active site;

lambda- - lambda lysogen deletion;

e14- absence of the prophage like element containing mcrA gene;

mcrA0 - : the McrA system is involved in the restriction of DNA sequences containing methylated cytosine at particular sequences;

rpsL150(strR) – mutation that confers streptomycin resistance;

mcrB1 - mutation eliminating restriction of DNA methylated at the sequence $R^{m}C$;

hsdR2 - for efficient transformation of cloned unmethylated DNA from PCR amplifications.

Supplementary Table 6.2 – List of plasmids used in the current work, with indication of size, general phenotype and reference.

Plasmid	Size (bp)	Description/Phenotype	Used in	Reference
pSM128	9086	Transcriptional phusion vector, SmR	Cloning of regulatory region	[63]
pJET1.2/blunt	2974	Linearized cloning vector, AmpR	Cloning of regulatory region	CloneJET PCR Cloning Kit (Thermo Fisher Scientific)
pYUB854	3893	Cosmid vector, with res sites flanking the Hygr gene	Phage-mediated mutagenesis	[79]
pHAE159	50725	Mycobacterial phage, AmpR	Phage-mediated mutagenesis	[79]
pYUB854_*pks1*	5732	Derived from pYUB854 by insertion of 2 regions of *pks1* flanking the Hygr gene	Phage-mediated mutagenesis / Mycobacterial recombineering	[80]
pJV53	8812	Vector containing an inducible acetamidase promoter controlling the recombination proteins gp60 and gp61, KanR	Mycobacterial recombineering	[68]

Supplementary Table 6.3 – List of primers used in the current work, with primer name, target, features, hybridization temperature (Tm) and reference.

Primer	Nucleotide sequence (5´- 3´)	Target	Location (bp)	Features	Tm (ºC)	Reference
*pks1*_F	ATCGAGTACTTGCATGTGGATGAGCCTTC	409 bp at *pks1* 5´region	3296573	ATCG with adaptor *ScaI* site	65.9	This work
*pks1*_R	ATCGAGTACTATCAGTTGCTCACGGCTTG		3296184		65.7	This work
*pks15*_F	ATCGAGTACTCGGGATTAGCCCAGTATTTG	508 bp at *pks15* 5´region	3298110	ATCG with adaptor *ScaI* site	64.8	This work
*pks15*_R	ATCGAGTACTGAAACGACATCCCAGAGTCC		3297623		64.5	This work
fadD22_F	ATCGAGTACTGCTAAAGTCTGGGTCGGAGA	336 bp at *fadD22* 5´region	3300167	ATCG with adaptor *ScaI* site	65.1	This work
fadD22_R	ATCGAGTACTGACCATGAGTCACCACATCG		3299852		64.9	This work

Supplementary Table 6.4 – List of accession codes used for synteny analyses of *pks1*, *pks15* and *fadD22*.

Accession code
*Mtb*_0A005DS_aa7362751
*Mtb*_0A012DS_aa7362551
*Mtb*_0A029DS_aa7362351
*Mtb*_0A033DS_aa7362151
*Mtb*_0A036DS_aa7361951
*Mtb*_0A087DS_aa7361751
*Mtb*_0A092DS_aa7361551
*Mtb*_0A093DS_aa7361351
*Mtb*_0A094DS_aa7361151
*Mtb*_0A115DS_aa7360951
*Mtb*_0A117DS_aa7360751
*Mtb*_0B026XDR_aa7360551
*Mtb*_0B049XDR_aa7360351
*Mtb*_0B070XDR_aa7360151
*Mtb*_0B076XDR_aa7359951
*Mtb*_0B123ND_aa7359751
*Mtb*_0B169XDR_aa7359551
*Mtb*_0B218DS_aa7359351
*Mtb*_0B222DS_aa7359151
*Mtb*_0B228DS_aa7358951
*Mtb*_0B229DS_aa7358751
*Mtb*_0B235DS_aa7358551
*Mtb*_0B259XDR_aa7358351
*Mtb*_0B329XDR_aa7358151
*Mtb*_1458_aa18552551
*Mtb*_1821ADB35_aa19974651
*Mtb*_1821ADB36_aa19974951
*Mtb*_1821ADB37_aa19975251
*Mtb*_1821ADB38_aa19975451
*Mtb*_1821ADB40_aa19975651
*Mtb*_1821ADB41_aa19975851
*Mtb*_1821ADB42_aa19976051
*Mtb*_1821ADB44_aa19976251
*Mtb*_1821ADB45_aa19976651
*Mtb*_18b_aa8351251
*Mtb*_22103_aa15450151
*Mtb*_22115_aa15449551
*Mtb*_2242_aa15448951
*Mtb*_2279_aa15449351
*Mtb*_26105_aa15450551
*Mtb*_37004_aa15449851
*Mtb*_49_02_aa7865051
*Mtb*_6A024XDR_aa7357951
*Mtb*_7199_99_aa3314451
*Mtb*_96075_aa7565251
*Mtb*_96121_aa7565451
*Mtb*_Beijing_391_aa21168151
*Mtb*_Beijing_aa17508651
*Mtb*_Beijing_like_1104_aa21168551
*Mtb*_Beijing_like_35049_aa21167551
*Mtb*_Beijing_like_36918_aa21167751
*Mtb*_Beijing_like_38774_aa21167951
*Mtb*_Beijing_like_50148_aa21168351
*Mtb*_Beijing_like_aa9541551
*Mtb*_BT1_aa5721751
*Mtb*_BT2_aa5721551
*Mtb*_C3_aa22901651
*Mtb*_CAS_NITR204_aa3899251
*Mtb*_CCDC5079_aa2703451
*Mtb*_CCDC5079_aa4006151
*Mtb*_CCDC5180_aa2703651
*Mtb*_CCDC5180_aa5721951
*Mtb*_CDC1551_aa85851
*Mtb*_CSV10399_aa24469351
*Mtb*_CSV11678_aa24469551
*Mtb*_CSV3611_aa24474151
*Mtb*_CSV383_aa24473951
*Mtb*_CSV4519_aa24468751
*Mtb*_CSV4644_aa24468951
*Mtb*_CSV5769_aa24469151
*Mtb*_CSV9577_aa24474351
*Mtb*_CTRI_2_aa2244351
*Mtb*_DK9897_aa19224851
*Mtb*_EAI5_aa4221251
*Mtb*_EAI5_NITR206_aa3899451
*Mtb*_F1_aa15446751
*Mtb*_F11_aa169251
*Mtb*_F28_aa15447051
*Mtb*_H37Ra_aa161451
*Mtb*_H37Ra_aa19387251
*Mtb*_H37Rv_aa1959552

*Mtb*_H37Rv_aa2777352

*Mtb*_H37Rv_aa8312451

*Mtb*_H37RvSiena_aa8270851

*Mtb*_HKBS1_aa5721251

*Mtb*_HN_024_aa23562551

*Mtb*_HN_205_aa23579351

*Mtb*_HN_321_aa23579551

*Mtb*_HN_506_aa23579751

*Mtb*_I0002353_6_aa18958451

*Mtb*_I0002801_4_aa18958251

*Mtb*_I0004000_1_aa18958051

*Mtb*_I0004241_1_aa18957651

*Mtb*_K_aa6984751

*Mtb*_KIT87190_aa7066651

*Mtb*_KZN_1435_aa236251

*Mtb*_KZN_4207_aa1545852

*Mtb*_KZN_605_aa1546052

*Mtb*_LE103_aa24475351

*Mtb*_LE13_aa24474551

*Mtb*_LE371_aa24475551

*Mtb*_LE410_aa24475751

*Mtb*_LE486_aa24469751

*Mtb*_LE492_aa24469951

*Mtb*_LE63_aa24474751

*Mtb*_LE76_aa24474951

*Mtb*_LE79_aa24475151

*Mtb*_LN1100_aa24477951

*Mtb*_LN180_aa24470151

*Mtb*_LN1856_aa24478151

*Mtb*_LN2358_aa24470351

*Mtb*_LN2900_aa24478351

*Mtb*_LN2978_aa24476551

*Mtb*_LN317_aa24476151

*Mtb*_LN3584_aa24476751

*Mtb*_LN3588_aa24476951

*Mtb*_LN3589_aa24477151

*Mtb*_LN3668_aa24477351

*Mtb*_LN3672_aa24477551

*Mtb*_LN3695_aa24477751

*Mtb*_LN3756_aa24470551

*Mtb*_LN55_aa24475951

*Mtb*_LN763_aa24476351

*Mtb*_M0002959_6_aa18958651

*Mtb*_M0018684_2_aa18957851

*Mtb*_MDRDM1098_aa24471151

*Mtb*_MDRDM260_aa24470751

*Mtb*_MDRDM627_aa24470951

*Mtb*_MDRDM827_aa24478551

*Mtb*_MDRMA1565_aa24479351

*Mtb*_MDRMA2019_aa24479551

*Mtb*_MDRMA203_aa24478751

*Mtb*_MDRMA2082_aa24479751

*Mtb*_MDRMA2260_aa24479951

*Mtb*_MDRMA2441_aa24480151

*Mtb*_MDRMA2491_aa24471351

*Mtb*_MDRMA701_aa24478951

*Mtb*_MDRMA863_aa24479151

*Mtb*_ME1473_aa24471551

*Mtb*_MTB1_aa20727751

*Mtb*_MTB1_aa20727752

*Mtb*_MTB2_aa22082351

*Mtb*_NCGM946K2_aa23560151

*Mtb*_PR08_aa9345851

*Mtb*_PR10_aa15847451

*Mtb*_RGTB327_aa2770851

*Mtb*_RGTB423_aa2771051

*Mtb*_S3_aa22901451

*Mtb_SCAI*D_187_0_aa12755652

*Mtb_SCAI*D_252_0_aa17082651

*Mtb_SCAI*D_320_0_aa17024351

*Mtb*_SLM036_aa24472551

*Mtb*_SLM040_aa24472751

*Mtb*_SLM056_aa24472951

*Mtb*_SLM060_aa24473151

*Mtb*_SLM063_aa24473351

*Mtb*_SLM088_aa24473551

*Mtb*_SLM100_aa24473751

*Mtb*_str_Beijing_NITR203_aa3648251

*Mtb*_str_Erdman_ATCC_35801_Erdman_ATCC35801_aa3502051

*Mtb*_str_Haarlem_aa1536852

*Mtb*_str_Haarlem_NITR202_aa3899051

*Mtb*_str_Kurono_aa8289951

*Mtb*_TB282_aa18701451

*Mtb*_TBDM1506_aa24480351

*Mtb*_TBDM2189_aa24480551

*Mtb*_TBDM2444_aa24480751

*Mtb*_TBDM2487_aa24480951

*Mtb*_TBDM2489_aa24481151

*Mtb*_TBDM2699_aa24481351

*Mtb*_TBDM2717_aa24481551	*Mtb*_TRS20_aa19978051
*Mtb*_TBDM425_aa24471751	*Mtb*_TRS21_aa19979351
*Mtb*_TBV4766_aa24481751	*Mtb*_TRS22_aa20009651
*Mtb*_TBV4768_aa24481951	*Mtb*_TRS23_aa19981051
*Mtb*_TBV4952_aa24482151	*Mtb*_TRS24_aa19987051
*Mtb*_TBV5000_aa24471951	*Mtb*_TRS25_aa20009051
*Mtb*_TBV5362_aa24472151	*Mtb*_TRS26_aa19987351
*Mtb*_TBV5365_aa24472351	*Mtb*_TRS27_aa19986451
*Mtb*_TRS1_aa19976951	*Mtb*_TRS28_aa19984851
*Mtb*_TRS10_aa20009251	*Mtb*_TRS29_aa20009451
*Mtb*_TRS11_aa19985851	*Mtb*_TRS4_aa19984551
*Mtb*_TRS12_aa19981751	*Mtb*_TRS5_aa19980751
*Mtb*_TRS13_aa19978851	*Mtb*_TRS6_aa19984251
*Mtb*_TRS14_aa19981451	*Mtb*_TRS7_aa19980251
*Mtb*_TRS15_aa19982251	*Mtb*_TRS8_aa19985451
*Mtb*_TRS16_aa19977251	*Mtb*_TRS9_aa19974251
*Mtb*_TRS17_aa19977751	*Mtb*_UT205_aa3045551
*Mtb*_TRS18_aa19982651	*Mtb*_W_148_aa1931852
*Mtb*_TRS19_aa19979751	*Mtb*_ZMC13_264_aa7384451
*Mtb*_TRS2_aa19983651	*Mtb*_ZMC13_88_aa7384751

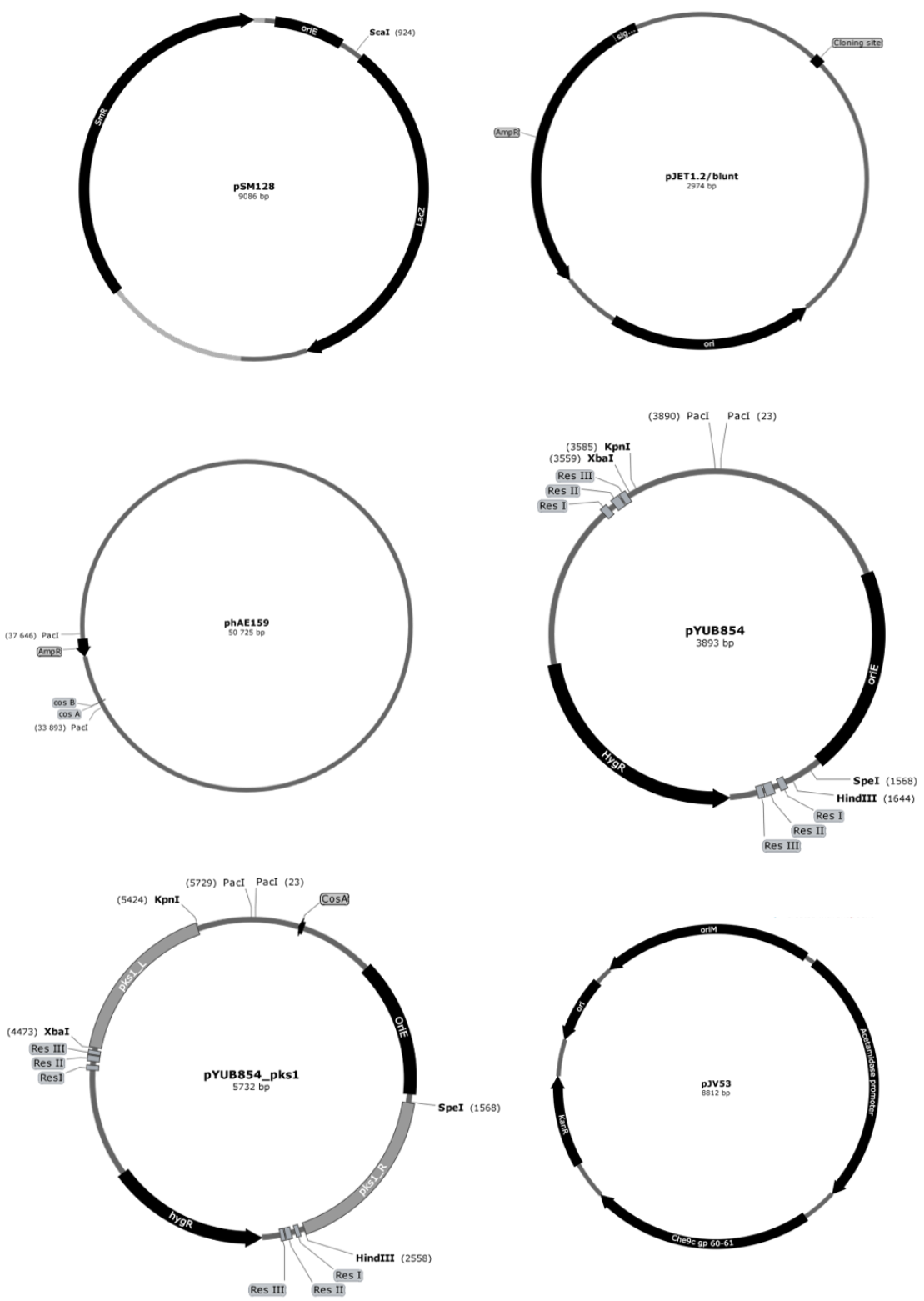

Figure 6.1 – Schematic representation of plasmids used in this work, with indication of restriction sites, antibiotic resistance cassettes and other relevant features.