**Guilherme Menegói Ribeiro**

# *Análise genômica de Mycobacterium massiliense GO 06*

Universidade de Brasília

Brasília, 19 de março de 2014

**Guilherme Menegói Ribeiro**

# *Análise genômica de Mycobacterium massiliense GO 06*

Dissertação de Mestrado apresentada como requisito parcial à obtenção do título de Mestre em Biologia Molecular.

Orientador:
Prof. Dr. Marcelo de Macedo Brígido

Co-orientador:
Drª. Tainá Raiol de Alencar

Universidade de Brasília
Instituto de Ciências Biológicas
Departamento de Biologia Celular
Programa de Pós-Graduação em Biologia Molecular

Universidade de Brasília

Brasília, 19 de março de 2014

# *Dedicatória*

Dedico este trabalho a todas as pessoas cujas vidas foram impactadas por infecções causadas por micobactérias não tuberculosas. Que esta pesquisa seja útil para que cada vez menos pessoas precisem passar pelas mesmas dificuldades.

# *Agradecimentos*

Gostaria de agradecer, primeiramente, à minha família: meus pais, Amilcar e Heloísa, e à minha irmã Vanessa. Acho que nenhum de vocês entende direito a natureza do meu trabalho, mas, mesmo assim, nunca houve falta de apoio financeiro (para os meus estudos) ou emocional em minha casa. Pai, mesmo sabendo que você não concorda com a minha escolha profissional, obrigado por me dar a liberdade de fazê-la, e de confiar que eram escolhas maduras e que eu estava preparado para lidar com suas consequências. Mãe, eu tenho certeza que você nunca entendeu direito o que um bioinformata faz, mas isso nunca te impediu de contar para todos os seus amigos, com o maior orgulho, que seu filho, aos 22 anos, finalizava um mestrado em bioinformática. Obrigado, mãe coruja!

Ao meu grupo de amigos, carinhosamente apelidado de *Brotherhood*, me parece até um pouco injusto não mencionar vocês juntamente com a minha família, afinal, eu conheço a maioria de vocês desde os 10 anos de idade, meus queridos irmãos que não compartilham meu sangue. Nós crescemos juntos e eu tenho certeza que tudo que eu sou hoje é um bom reflexo de nossa convivência e das experiências que compartilhamos. Dessa forma, acredito que se esta tese é uma vitória minha, ela também é uma pequena vitória de vocês.

Também gostaria de agradecer à Fernanda, minha madrinha de curso e amiga de longa data, e ao Marco e à Aline, que me acompanharam durante toda a graduação, e hoje, assim como eu, são mestrandos na Universidade de Brasília. Como companheiros biólogos, vocês, mais do que ninguém, conseguem entender as maravilhas e as frustrações da pesquisa científica. Obrigado pelos almoços e lanchinhos da tarde, esses momentos eram muito especiais para esquecer aquele experimento do dia que não deu certo!

Aos que tornaram possível este trabalho: Prof. Marcelo Brígido, obrigado pela confiança em me acolher como orientando e pela grande contribuição que seu vasto conhecimento biológico trouxe ao meu trabalho. À Tainá, umas das pessoas mais solícitas e pacientes que eu já conheci na minha vida. Sou grato por toda a sua ajuda e orientação neste trabalho e compreensão nos momentos em que não pude me doar a ele da maneira como deveria, e posso dizer, sem sombra de dúvidas, que devo à você a maior parte do meu conhecimento em bioinformática. Aos professores André Kipnis e Ana Paula

Junqueira-Kipnis, que trabalham com *M. massiliense* e cujos esforços foram determinantes na realização deste trabalho. À Prof. María Emília, que me ensinou tudo que eu sei de algoritmos computacionais de bioinformática e cujas sugestões foram essenciais para a finalização deste trabalho. Um agradecimento especial também ao meu colega João Araújo, que ajudou este simples biólogo quando ele encontrava um problema computacional insolucionável!

Também não poderia deixar de agradecer aos professores Nalvo e Chyntia, que tão prontamente aceitaram o convite para participar da banca da minha defesa. Suas sugestões foram muito importantes para a excelência desta dissertação. Não me esqueço também de agradecer à Ana, da secretaria de Pós-Graduação, sempre muito solicita no auxílio com os problemas burocráticos do mestrado. Obrigado!

Finalmente, mas não menos importante, gostaria de agradecer especialmente ao meu companheiro, Lucas. Obrigado por me suportar nos momentos em que eu não tinha mais paciência ou inspiração para escrever esta tese. Obrigado por todas as noites em claro me ajudando a corrigir pequenos erros de português, ou reescrevendo frases para que ficassem mais didáticas. Obrigado, acima de tudo, por escolher compartilhar a sua vida comigo: nos conhecemos há apenas um ano e meio, e eu acho que já aprendi mais com você do que 10 mestrados poderiam me ensinar.

# *Resumo*

As micobactérias de crescimento rápido (RGM) têm implicações importantes em patologias humanas, sendo relacionadas à infecções oportunistas. Desde sua descrição em 2004, os casos clínicos associados a *Mycobacterium massiliense*, uma espécie representativa do grupo das RGM, tem sido crescentemente reportados. Com o aumento dos surtos causados por essas bactérias, o desenvolvimento de novas técnicas de detecção de espécie e de predição de padrões de susceptibilidade é essencial para o controle eficiente de suas infecções. O objetivo deste trabalho é utilizar a genômica comparativa para traçar um perfil do funcionamento biológico e dos mecanismos de virulência de *M. massiliense*, facilitando a descoberta de moléculas de interesse. A estirpe GO 06 de *M. massiliense* foi isolada durante o surto ocorrido no período entre 2005 e 2007 em Goiás, na região central do Brasil, e teve seu genoma sequenciado pela plataforma de sequenciamento de alto desempenho 454 GS-FLX Titanium (Roche). Foi possível montar o genoma completo da estirpe GO 06, constituído por seu cromossomo e dois plasmídos, bem como anotar a maioria das suas ORFs preditas (3.491 ORFs, representando 84,5% do total identificado), com a identificação de 826 genes relacionados à virulência. Também foi possível identificar 46 tRNAs e um único operon de rRNA. As vias metabólicas relacionadas aos sistemas de secreção bacterianos e à bioissíntese de sideróforos de *M. massiliense* GO 06, geralmente envolvidas na patogenicidade micobacteriana, foram descritas *in silico*. 15 genes relacionados ao T7SS também foram identificados no genoma do isolado GO 06, sugerindo a existência dos sistemas ESX-3 e ESX-4. Os dados gerados neste projeto fornecem informações importantes para o desenvolvimento de estratégias de controle de surtos relacionados às RGM, sendo disponibilizados em uma página hospedada no domínio público da Universidade de Brasília.

Palavras-chave: *Mycobacterium massiliense*; genoma; bioinformática; fatores de virulência; sideróforos; sistemas de secreção bacterianos; sistema de secreção do tipo 7; T7SS;

# *Abstract*

Rapid growing mycobacteria (RGM) have important implications in human diseases, being often related to opportunistic infections. Since its description in 2004, clinical cases related to *Mycobacterium massiliense*, a representative species of the RGM group, have been increasingly reported. With the increase of outbreaks related to these bacteria, the development of new species detection and susceptibility pattern prediction techniques is essential for the efficient control of their infections. The goal of this project is to use comparative genomics to trace the biological functioning and virulence mechanisms profiles of *M. massiliense*, facilitating the discovery of molecules of interest. The strain GO 06 of *M. massiliense* was isolated during the outbreak that occurred between 2005 and 2007 in Goiás, in the midwest region of Brazil, and had its entire genome sequenced using the 454 GS-FLX Titanium (Roche) high-throughput sequencer. It was possible to construct strain GO 06's entire genome, which was comprised of its chromosome and two plasmids, and annotate the majority of its predicted ORFs (3.491 ORFs, 84,5% of all identified ORFs), with the identification of 826 genes related to virulence. It was also possible to identify 46 tRNAs and a single rRNA operon. *M. massiliense* GO 06's metabolic pathways regarding bacterial secretion systems and siderophore biosynthesis, usually involved in mycobacterial pathogenicity, were described *in silico*. 15 genes related to T7SS were also identified in isolate GO 06's genome, suggesting the existence of ESX-3 and ESX-4 systems. The data generated in this project represents useful information in the development of control strategies of outbreaks related to RGM, being made available in a webpage hosted in the public domain of University of Brasília.

Keywords: *Mycobacterium massiliense*; genome; bioinformatics; virulence factors; siderophores; bacterial secretion systems; type 7 secretion system; T7SS;

# *Sumário*

# *Lista de Figuras*

# *Lista de Tabelas*

# *Lista de Símbolos, Siglas e Abreviaturas*

**APS** *Adenosine 5'-phosphosulfate*, adenosina 5'-fosfosulfato

**BLAST** *Basic Local Alignment Search Tool*

**COG** *Clusters of Orthlogous Groups*

**DNA** *Deoxyribonucleic acid*, ácido desoxirribonucleico

**dNTP** *Deoxyribonucleotide triphosphate*, desoxirribonucleotídeo trifosfato

**ESX** *ESAT-6 secretion system*, sistema de secreção de ESAT-6

**GRC** *Genome Reverse Compiler*

**ITS** *Internal Transcribed Spacer*

**KEGG** *Kyoto Encyclopedia of Genomes and Genes*

**LAM** Lipoarabinomanano

**LPS** Lipopolissacarídeo

**NCBI** *National Center for Biotechnology Information*

**NTM** *Nontuberculous mycobacteria*, micobactéria não tuberculosa

**OLC** *Overlap-layout-consensus*

**PCR** *Polymerase chain reaction*, reação em cadeia da polimerase

**RGM** *Rapid growing mycobacteria*, micobactéria de crescimento rápido

**RNA** *Ribonucleic acid*, ácido ribonucleico

**rRNA** RNA ribossomal

**SGM** *Slow growing mycobacteria*, micobactéria de crescimento lento

**T1SS** *Type 1 secretion system*, sistema de secreção do tipo 1

**T3SS** *Type 3 secretion system*, sistema de secreção do tipo 3

**T7SS** *Type 7 secretion system*, sistema de secreção do tipo 7

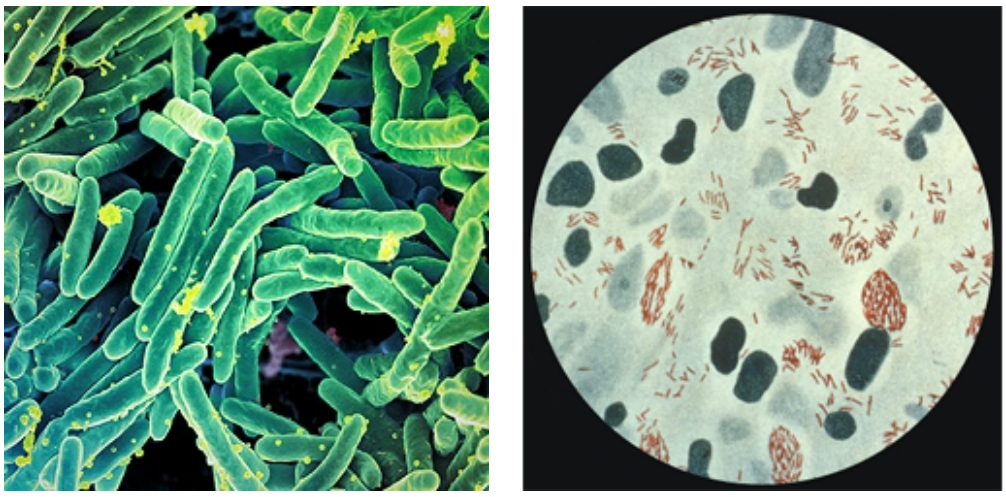
**tRNA** RNA transportador

**VFDB** *Virulence Factor Database*

# 1 Introdução

## 1.1 Micobactérias

Micobactérias são bactérias pertencentes ao gênero *Mycobacterium* do filo Actinobacteria, um dos gêneros bacterianos mais estudados devido à sua associação com diversas doenças infecciosas. Seus representantes mais conhecidos são os agentes etiológicos da tuberculose, *Mycobacterium tuberculosis* (Figura 1a), e da hanseníase, *Mycobacterium leprae* (Figura 1b).



(a) *Mycobacterium tuberculosis* (b) *Mycobacterium leprae*

Figura 1: **(a)** Microscopia eletrônica de varredura de *M. tuberculosis*. Fonte: `http://www.niaid.nih.gov/LabsAndResources/resources/translational/microscopy/Pages/sem.aspx`. **(b)** Microscopia ótica de uma lesão cutânea causada por *M. leprae*. As bactérias estão coradas em vermelho. Fonte: `http://www.ppdictionary.com/bacteria/gpbac/leprae.htm`.

As micobactérias são bacilos imóveis, aeróbios obrigatórios e não formadores de esporos, que em determinada etapa do seu ciclo de vida apresentam a característica de álcool-ácido resistência [3]. Essa propriedade é proveniente da presença, na superfície da parede celular dessas bactérias, de ácidos graxos ramificados de cadeia longa denominados "ácidos micólicos", e é caracterizada pela resistência à descolorização por ácidos [4]. Como consequência, a técnica de Gram, corriqueiramente utilizada para a separação das

bactérias em dois grupos distintos (Gram-Positivas e Gram-Negativas), tem pouco efeito na identificação das micobactérias.

Apesar de não serem coradas pela técnica de Gram, micobactérias são comumente classificadas como Gram-positivas, já que não apresentam uma membrana externa em sua parede celular, similarmente a outras bactérias desta classificação. A camada de ácido micólico (também chamada de micomembrana) é ancorada aos peptideoglicanos por moléculas de arabinogalactano. Outro componente de importância são os lipoarabinomananos (LAM), moléculas que se estendem da membrana plasmática à superfície celular, estruturalmente e funcionalmente análogas aos lipopolissacarídeos (LPS) de paredes celulares de bactérias Gram-negativas [5]. A Figura 2 mostra uma representação de parede celular micobacteriana.

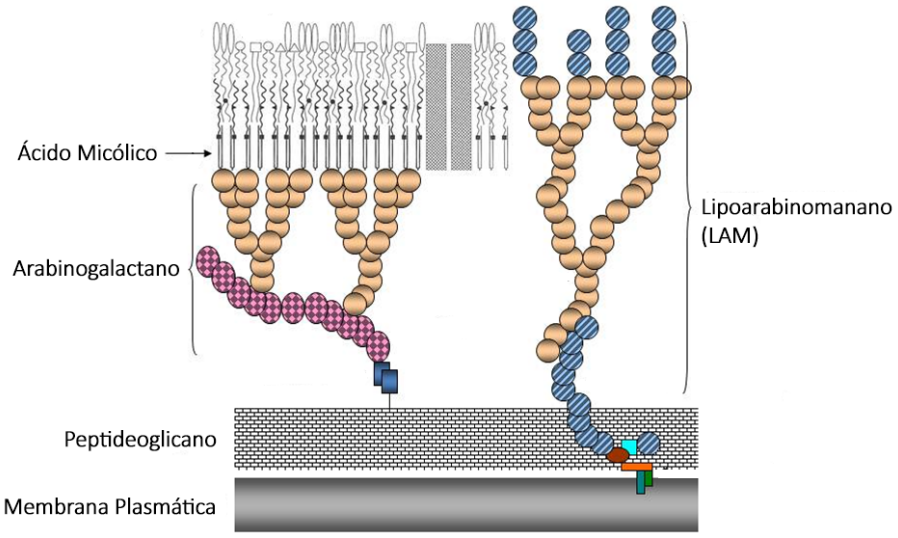



Figura 2: Representação esquemática da parede celular de uma micobactéria. Fonte: adaptado de `http://www.utoronto.ca/liulab/Projects.html`.

O alto teor de lipídios da parede celular atua como uma camada cerosa que a deixa mais resistente e hidrofóbica, e é particularmente responsável por grande parte das características biológicas dessas bactérias. Além da influência na supracitada propriedade de resistência ao álcool-ácido, a parede menos permeável também garante a essas bactérias maior tolerância a vários antibióticos [7]. Por fim, a hidrofobicidade da parede ainda possibilita maior adesão entre indivíduos em uma colônia [3].

Em geral, as micobactérias podem ser separadas em dois grupos distintos: as micobactérias de crescimento lento (do inglês, *slow growing mycobacteria*, ou SGM) e as de crescimento rápido (do inglês, *rapid growing mycobacteria*, ou RGM). Um exemplo típico de SGM são as micobactérias do complexo *M. tuberculosis*, e micobactérias que não fa-

zem parte deste grupo são comumente denominadas "micobactérias não tuberculosas" (do inglês, *nontuberculous mycobacteria*, ou NTM), sendo de crescimento rápido ou não.

### 1.1.1 Micobactérias de crescimento rápido (RGM)

As micobactérias de crescimento rápido são aquelas que, por definição, formam colônias em até sete dias de incubação [8]. Pertencem a este grupo as micobactérias do complexo *Mycobacterium fortuitum* e do complexo *Mycobacterium chelonae-Mycobacterium abscessus*. As RGM são ubíquas, havendo relatos do seu isolamento a partir de amostras de solos, rochas, e de água [9]. Alguns estudos demonstram ainda que, devido à parede celular hidrofóbica, essas bactérias conseguem sobreviver a ambientes extremos formando biofilmes bastante agregados [10].

Ao contrário das bactérias do complexo *M. tuberculosis* ou de *M. leprae*, não existem evidências de que as infecções por RGM sejam transmissíveis entre humanos [11]. As infecções normalmente se dão pela inalação ou ingestão de bacilos viáveis, ou pela introdução dos mesmos em lesões cutâneas. Por conta de sua ubiquidade, as micobactérias de crescimento rápido são associadas a infecções oportunistas em pacientes que sofreram intervenções cirúrgicas ou são imunocomprometidos [9, 12].

Os sintomas das infecções causadas por RGM são variados e pouco aplicáveis na identificação de espécies [8]. Essa classe de micobactérias apresenta, ainda, resistência a diversos antibióticos e desinfetantes [13], sendo resistente, inclusive, a drogas antituberculosas como Rifampicina e Isoniazida [14]. A grande variedade de RGM com sintomas similares e sua resistência a fármacos dificulta o tratamento de infecções causadas por esses organismos.

### 1.1.2 *Mycobacterium massiliense*

Em 2004, Adékambi *et al.* [16] descreveram uma nova espécie de RGM intimamente relacionada a *M. abscessus*, isolada de um paciente com pneumonia hemoptoica. A espécie foi caracterizada por métodos bioquímicos e moleculares, entre eles testes de características fenotípicas e de susceptibilidade a antibióticos, análise do conteúdo de GC do DNA e análise genotípica de diversos genes. A nova espécie, nomeada *Mycobacterium massiliense*, se diferenciava de *M. abscessus* por divergências nos genes *rpoB*, *recA*, *hsp65*, *sodA* e na região ITS 16S-23S rRNA; e por algumas características fenotípicas que se encontram resumidas na Tabela 1.

Tabela 1: Características fenotípicas que diferenciaram as espécies *M. abscessus* e *M. massiliense*, como descrito por Adékambi *et al.* [16]. As características presentes em uma determinada bactéria são representadas por um "+" e as ausentes por um "-".

| | *Mycobacterium abscessus* | *Mycobacterium massiliense* |
|---|---|---|
| ß-galactosidase | - | + |
| N-acetil-ß-glucosaminidase | - | + |
| ß-glucuronidase | + | - |
| Nitrato redutase | + | - |
| Produção de indol | + | - |
| Resistência a doxiciclina | + | - |

Desde a sua descrição, porém, a classificação de *M. massiliense* tem sido bastante controversa, em particular pela dificuldade em atribuir uma espécie a micobactérias muito próximas. A sequência do gene 16S rRNA, por exemplo, é idêntica entre todas as micobactérias do complexo *M. abscessus* (*M. abscessus*, *M. massiliense* e *Mycobacterium bolletii*), o que impossibilita a discriminação entre essas três espécies por meio da comparação deste gene, uma técnica rotineira em muitos laboratórios [8]. Muitos estudos propõem que esta diferenciação seja feita por outros marcadores genéticos, porém os resultados desses estudos se mostraram conflitantes [17, 18].

Algumas características fenotípicas se mostraram igualmente inapropriadas para essa caracterização. Diferentemente da suscetibilidade inicial a doxiciclina reportada, por exemplo, um estudo identificou cepas de *M. massiliense* resistentes a este antibiótico [19]. Além disso, outros estudos demonstraram que algumas cepas de *M. massiliense* têm atividade de urease ou suscetibilidade a claritromicina, o que também é discrepante em relação à descrição inicial da espécie [18].

Apesar da dificuldade na classificação desses organismos, vale ressaltar que a maioria dos estudos que buscam identificar espécies de RGM por meio de marcadores genéticos se concentram na comparação de um número pequeno de genes (a maioria deles apenas analisa o *rpoB*). Estudos multilocus (isto é, que levam em consideração vários marcadores) se mostraram mais eficazes na discriminação das espécies do complexo *M. abscessus* [17, 20].

### 1.1.3 Fatores de virulência

Bactérias patogênicas precisam ser capazes de invadir seu hospedeiro e de sobreviver e se replicar enquanto evitam os mecanismos de proteção do organismo invadido. Logo, elas apresentam um conjunto de mecanismos moleculares que as tornam capazes de evitar o sistema imune do hospedeiro, comumente chamados de "fatores de virulência". O estudo das moléculas dessa classificação pode contribuir na descoberta de novos agentes terapêuticos. Alguns fatores de virulência comuns em micobactérias estão resumidos na Tabela 2.

Tabela 2: Fatores de virulência comumente encontrados em micobactérias. Informações obtidas de [21].

| Fator de virulência | Função biológica |
|---|---|
| Micobactina | Captação de ferro |
| LAM | Componente de parede |
| Antígeno 85 | Componente de parede |
| Isocitrato liase | Metabolismo celular |
| ESAT-6/CFP-10 | Proteína secretada |
| ESX-1 | Sistema de secreção |
| ESX-5 | Sistema de secreção |
| sodA | Proteína de estresse |
| Fosfolipase C | Toxina |

Em *M. tuberculosis*, conhece-se o papel dos sistemas de captação de ferro [21, 22] e do envelope celular [23, 24] na patogenicidade desse organismo. O ferro é um micronutriente essencial para o crescimento bacteriano, e a falta deste nutriente pode induzir a parada do crescimento de *M. tuberculosis* [22]. Apesar de sua importância, o ferro é um dos nutrientes de menor solubilidade em água, sendo virtualmente insolúvel em faixas de pH neutro. Dessa forma, sua acessibilidade somente é viável por meio de mecanismos de captação de ferro: moléculas quelantes ou transportadoras de íons. Em bactérias, as moléculas responsáveis por essa captação, seja competindo com o hospedeiro por seus estoques de ferro ou adquirindo-o diretamente do ambiente, são chamadas "sideróforos". Em geral, as micobactérias produzem dois sideróforos: uma molécula intracelular, a micobactina, e uma molécula extracelular, a carboximicobactina (em patógenos) ou a exoquelina (em bactérias saprofíticas) [21].

Já o envelope celular cria um microambiente de permeabilidade reduzida a moléculas polares e de grande tamanho, como alguns antibióticos [7, 23]. Além disso, vários compo-

nentes do envelope celular têm funções biológicas relevantes, que se interrompidas podem afetar a viabilidade da bactéria. Já foi demonstrado, por exemplo, que o LAM é um componente essencial para a integridade da parede celular micobacteriana, e que sua ausência aumenta a suscetibilidade a antibióticos e à fagocitose [24].

Outro fator de virulência de interesse são os sistemas de secreção, mecanismos especializados na secreção de proteínas bacterianas para o meio extracelular. Além dos sistemas Sec-SRP e Tat, que funcionam como sistemas gerais de secreção e são comuns tanto em bactérias Gram-positivas quanto em Gram-negativas, existem ainda os sistemas de secreção alternativos, que encaminham proteínas bacterianas para a célula do hospedeiro e manipulam a resposta do hospedeiro à infecção [25]. Estes sistemas são bastante comuns em bactérias Gram-negativas, já que a maior complexidade e menor permeabilidade de sua parede celular fazem necessária a presença de tal especialização. Existem descritos seis tipos de sistema de secreção alternativos em bactérias Gram-negativas (tipo I ao VI) [26]. Apesar de serem classificadas como Gram-positivas, a parede celular de micobactérias é mais complexa devido à presença dos ácidos micólicos, e a presença de um sistema similar ao sistemas de secreção alternativos já havia sido predita por análises *in silico* [27]. Na última década, estudos demonstraram não só a presença de um mecanismo de secreção alternativo em micobactérias - o sistema de ESX-1, mais tarde nomeado de "sistema de secreção do tipo VII" -, mas também correlacionaram a deleção deste sistema com a perda da virulência por *M. tuberculosis* [28, 29]. No genoma deste organismo foram identificados ainda outros quatro *loci* com genes homólogos aos do ESX-1, nomeados ESX-2 a ESX-5 [26].

### 1.1.4 Epidemiologia

Com o avanço das técnicas de cultura e de identificação, o número de casos reportados relacionados a NTM tem aumentado nos últimos anos. Bactérias representativas do grupo das RGM, particularmente as micobactérias do complexo *M. chelonae-M. abscessus* emergiram como uma das principais causadoras de doenças entre as NTM [8, 9, 18, 30]. Aproximadamente 65 a 80% dos casos de doença pulmonar causados por NTM, por exemplo, tem como agente etiológico *M. abscessus* [31].

Desde sua descrição, o número de casos relatados de infecções por *M. massiliense* também tem aumentado [19, 20]. No Brasil, surtos de infecções causadas por *M. massiliense* têm sido reportados crescentemente [18]. Entre 2004 e 2005, um surto que acometeu 311 pacientes submetidos a intervenções invasivas em Belém, na região norte do Brasil, foi

atribuido a *M. massiliense* e *M. bolletii* [32]. Na região Centro-Oeste do país, na cidade de Goiânia, 121 casos de infecção por NTM foram notificados ao sistema de saúde público no período de 2005 a 2007. Grande parte dos isolados clínicos foram identificados como *M. massiliense* [33]. Por fim, no período entre Agosto de 2006 e Julho de 2007, mais de mil casos de doenças relacionadas a RGM foram relatados na região Sudeste do país, no estado do Rio de Janeiro. A maioria dos isolados clínicos foram identificados como *M. massiliense* ou *M. bolletii* [34].

## 1.2 Análise de genomas bacterianos

Embora sequências provenientes de vários organismos estivessem disponíveis anteriormente, a genômica comparativa moderna só surgiu com o sequenciamento dos dois primeiros genomas bacterianos, *Haemophilus influenzae* [35] e *Mycoplasma genitalium* [36]. A comparação de genomas foi um marco importante, pois possibilitou o delineamento de conjuntos de genes ortólogos e a determinação de genes que estão ausentes em certos organismos [37].

A genômica comparativa tem aplicabilidade em diversas áreas do conhecimento: na descrição de vias metabólicas [38] e na engenharia metabólica [39, 40], no estudo da expressão de genes [41], na descrição de novos alvos terapêuticos [42], nos estudos da evolução da patogenicidade de bactérias [43], entre outros. Em *M. tuberculosis*, a genômica comparativa proporcionou um maior entendimento do funcionamento da bactéria [44], bem como a predição de sistemas de virulência [45] e alvos terapêuticos [46].

Com o avanço das técnicas de sequenciamento e o desenvolvimento de *pipelines* para montagem e anotação de genomas, o número de genomas bacterianos completos cresceu exponencialmente. Em setembro de 2012, de acordo com dados do Genomes Online Database [47], haviam 3.699 genomas finalizados e disponibilizados online. Em Janeiro de 2014 esse número aumentou para 12.724, sendo 12.095 genomas bacterianos. Hoje é possível, em relativamente pouco tempo, anotar um genoma suficientemente bem para delinear um esquema do funcionamento celular do organismo e suas principais vias metabólicas [37].

## 1.3 Análise de bioinformática

Atualmente, o volume de dados biológicos disponíveis em bancos de dados é gigantesco, e tende a aumentar à medida que a tecnologia de sequenciamento avança e os custos

associados a essa técnica diminuem (Figura 3). Por outro lado, os algoritmos utilizados na comparação de sequências são demasiado complexos para sua implementação manual. Seria inviável processar o grande volume de dados obtidos por sequenciadores de nova geração manualmente.

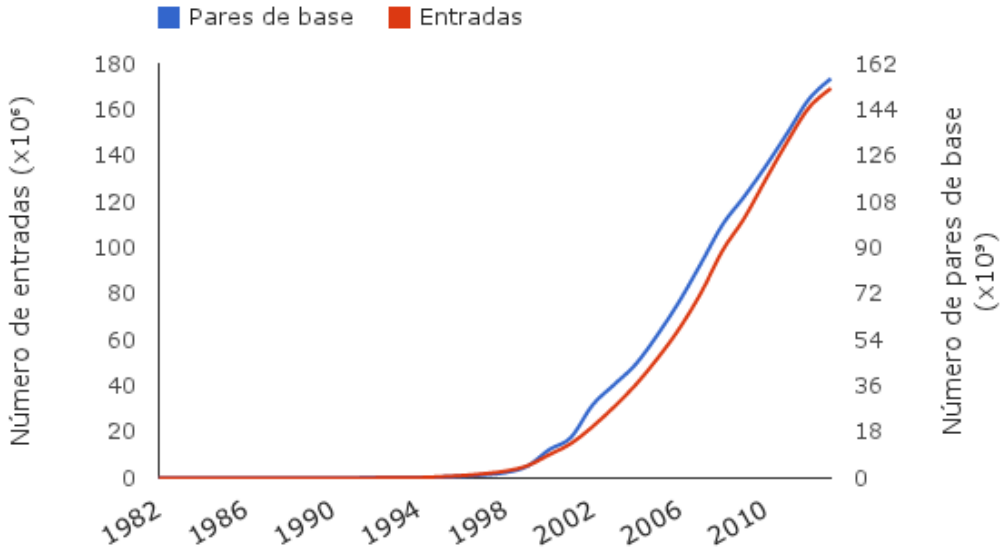



Figura 3: Número de entradas e volume total de sequências (por pares de base) depositadas no repositório do NCBI, desde a sua criação até 2013. Fonte: elaborada pelo autor a partir dados obtidos do NCBI [48].

A bioinformática é a ciência que busca facilitar o processamento de dados biológicos por meio da automação de processos e da manipulação destes dados por ferramentas e técnicas computacionais.

### 1.3.1 Técnicas de sequenciamento

O primeiro passo na análise genômica é o sequenciamento do genoma do organismo. Sequenciamento é o processo de determinação da estrutura primária de um biopolímero. No caso das moléculas de DNA, o objetivo é conhecer a sequência de nucleotídeos de um determinado fragmento.

Apesar de alguns fragmentos pequenos de DNA já terem sido sequenciados anteriormente, o sequenciamento moderno nasceu com a metodologia de terminação de cadeia proposta por Sanger em 1977 [49]. Uma grande desvantagem do sequenciamento Sanger é a necessidade da amplificação dos fragmentos de DNA a serem sequenciados [50]. Essa amplificação é geralmente feita através da clonagem em bactérias hospedeiras, um pro-

cesso passível de erros e bastante trabalhoso [51]. Outras desvantagens do sequenciamento Sanger são o baixo paralelismo, isto é, a quantidade de fragmentos que podem ser sequenciados ao mesmo tempo, e a necessidade da aplicação de eletroforese para identificação das bases. Com advento das plataformas de sequenciamento de próxima geração, tornou-se possível o sequenciamento em maior escala e a menor preço, assim como mostram os dados da Tabela 3.

Tabela 3: Comparação entre e as principais plataformas de sequenciamento de próxima geração. Dados obtidos de [52].

| | ABI 3730xl | Illumina HiSeq 2000 | 454 GS FLX Titanium | SOLiD 5500xl |
|---|---|---|---|---|
| Tamanho médio do fragmento sequenciado (pb) | 900 | 150 | 700 | 85 |
| Custo do sequenciamento (US$ por milhão de base) | 500 | 0,02 | 12,56 | 0,04 |
| Volume de bases sequenciadas por corrida | 2,88 Mb | 600 Gb | 0,7 Gb | 30 Gb |
| Tempo de corrida | 3 horas | 8 dias | 1 dia | 7 dias |

Apesar do custo reduzido, da possibilidade de sequenciar mais fragmentos em menos tempo e da necessidade de menos material biológico por corrida, a principal desvantagem dos sequenciadores de nova geração é a incapacidade de sequenciar fragmentos longos [50], o que prejudica o processo de montagem, principalmente em fragmentos de alta repetição.

Dentre as tecnologias de sequenciamento de próxima geração, o 454 é a que gera os maiores fragmentos sequenciados (*reads*), apesar do volume de bases sequenciadas ser menor que em outros sequenciadores de segunda geração. A utilização da plataforma 454 no sequenciamento de genomas bacterianos é bastante comum, já que, por serem genomas geralmente pequenos, não requerem grande volume de dados para a sua montagem.

Na tecnologia 454 [53], a amostra de DNA é fragmentada por sonicação, e em cada extremidade dos fragmentos originados são adicionados adaptadores. Esses adaptadores fixam os fragmentos a contas (*beads*) cobertas por estreptavidina, sendo que a estequiometria da reação é calculada de forma que cada *bead* contenha apenas um fragmento. As contas são emulsificadas em uma mistura de água com óleo, e cada conta é capturada em seu próprio microrreator, onde a reação de PCR para a amplificação dos fragmentos ocorre. As *beads* são então depositadas em uma placa com poços de aproximadamente

29 $\mu$m (*Picotiter plate*, ou PTP), espaço suficiente para a inserção de apenas uma *bead*. Após a centrifugação da placa para que ocorra o posicionamento das *beads*, as enzimas e reagentes necessários para o pirossequenciamento são depositados nos poços [50].

A reação de pirossequenciamento (Figura 4) baseia-se na detecção da incorporação de nucleotídeos por meio da emissão de luz. A cada ciclo do reator, um dos quatro deoxirribonucleotídeos trifosfato (dNTPs) é adicionado à reação, havendo a incorporação destes, se o pareamento de bases for correto, na fita nascente de DNA. O pirofosfato liberado durante esta incorporação é, na presença de APS, prontamente convertido em ATP pela enzima sulfurilase. A enzima luciferase, então, utiliza esse ATP na conversão de luciferina em oxiluciferina, um processo que leva à liberação de luz visível. A luz é captada pelo sequenciador, sendo proporcional à quantidade de dNTPs incorporados no ciclo, e a base incorporada é registrada [53].

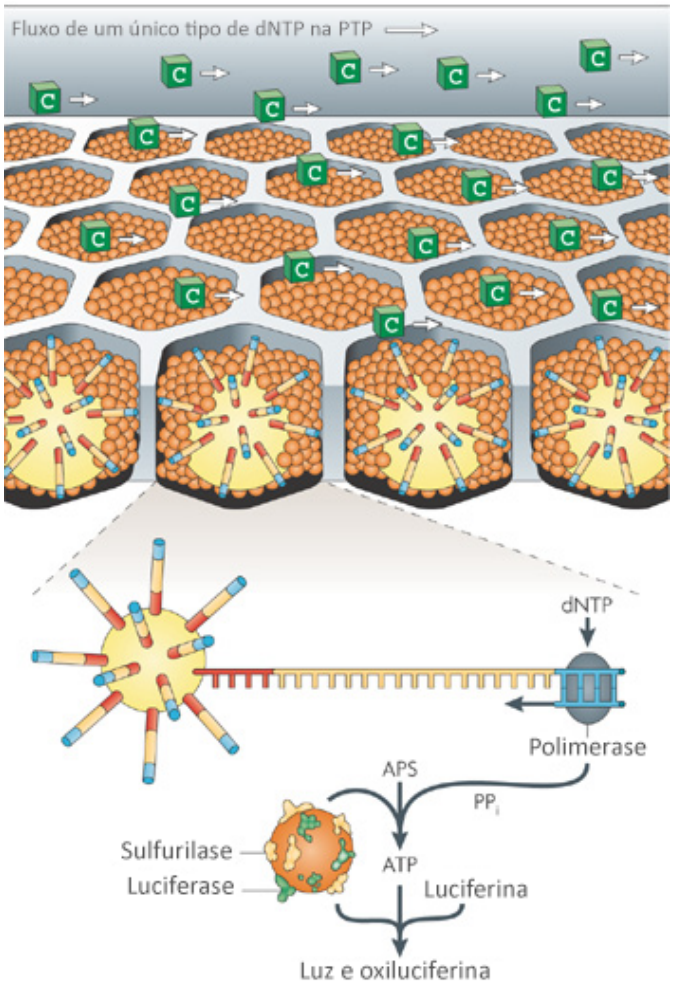



Figura 4: Representação esquemática da reação de pirossequenciamento. Fonte: adaptada de Metzeker, M.L. [54].

### 1.3.2 *Pipeline* computacional

Com o genoma sequenciado, o primeiro passo na análise por bioinformática é a elaboração de um *pipeline* computacional. Um *pipeline* é uma série de processos em sequência, onde a saída de um processo (*output*) e a entrada do próximo processo da série (*input*). A Figura 5 mostra um *pipeline* genérico normalmente empregado nas análises de genoma.

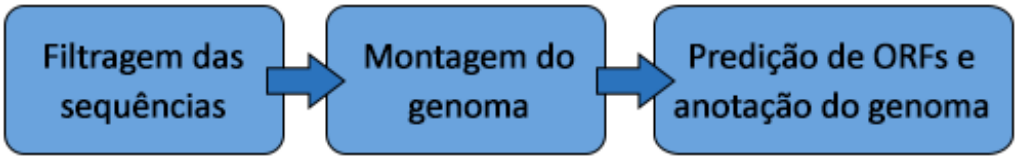



Figura 5: Representação esquemática de um *pipeline* genérico utilizado na análise de genomas por bioinformática. Fonte: elaborada pelo autor.

Na análise de genomas, o primeiro processo do *pipeline* de análise computacional é a filtragem dos fragmentos sequenciados e a análise de sua qualidade. Em seguida, esses fragmentos são montados em um genoma, que é então anotado.

### 1.3.3 Filtragem

Grande parte das sequências retiradas diretamente dos sequenciadores automáticos apresentam erros de incorporação de base [55]. Os procedimentos de filtragem buscam aumentar a confiabilidade dos dados obtidos do sequenciador, seja removendo das sequências os adaptadores de sequenciamento (*clipping*), ou fragmentos de baixa qualidade (*trimming*). Durante o *clipping* também ocorre a comparação das sequências com um banco de sequências provenientes de vetores e contaminantes comuns. Desta forma, fragmentos das *reads* similares às sequências do banco são removidos, evitando-se então o comprometimento dos dados por incorporação de fragmentos alienígenas [55].

A qualidade das sequências é avaliada pelo Phred *quality score* [56], uma escala de pontuação originalmente empregada pela ferramenta Phred, desenvolvida para a automação do sequenciamento do projeto Genoma Humano. A fórmula utilizada no cálculo do Phred *quality score* segue abaixo.

$$Q = -10 \log_{10} P$$

$Q =$ Phred *quality score*

$P =$ probabilidade da base ter sido incorporada erroneamente

Atualmente, a escala de pontuação Phred é aceita universalmente na análise da qualidade das sequências de DNA, sendo, inclusive, utilizada para comparação de sequenciadores diferentes. Um Phred *score* de pelo menos 20 é considerado adequado (o que significa que existe uma chance em cem de a base incorporada pelo sequenciador estar errada).

### 1.3.4 Montagem do genoma

Devido à limitação das técnicas de sequenciamento, que só permitem o sequenciamento direto de fragmentos de DNA relativamente curtos (30 a 900 nucleotídeos) [50], o DNA genômico precisa ser fragmentado antes de ser sequenciado. Logo, é necessário remontar os fragmentos obtidos do sequenciador na ordem correta. Esse processo de montagem é feito por *softwares* especializados, como o AMOS [57] ou o MIRA assembler [58].

A montagem de genomas é baseada na sobreposição de *reads*. Como o DNA genômico é amplificado e em seguida fragmentado aleatoriamente, espera-se que existam várias cópias de cada região genômica em fragmentos de DNA diferentes. Quando duas regiões de fragmentos diferentes se sobrepõem, elas são unidas em uma única sequência de DNA contígua, o *contig*, ou seja, é formada uma sequência consenso de tamanho maior (Figura 6). Existem dois principais métodos de montagem: o *de novo*, em que as *reads* são montadas *ab initio*; e o *mapping*, em que a montagem baseia-se em um genoma de referência, usualmente proveniente de um organismo relacionado.

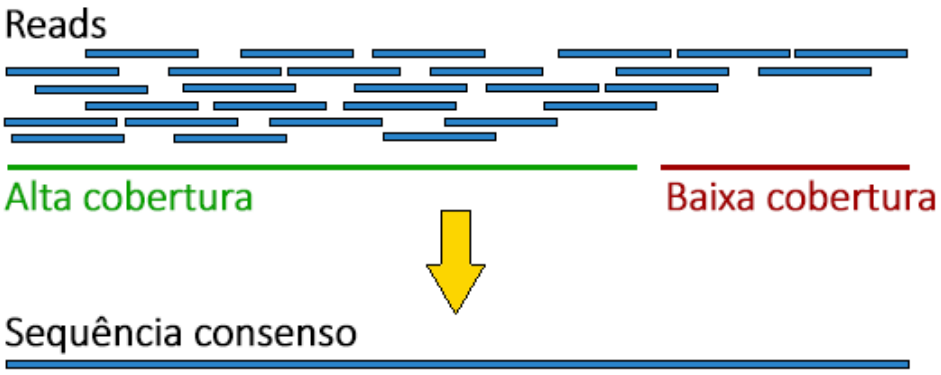



Figura 6: Representação esquemática da geração de uma sequência consenso (*contig*) a partir de reads. Fonte: adaptada de `http://gcat.davidson.edu/phast/`.

A qualidade de uma montagem normalmente é medida pela estatística N50. Para calcular o N50, ordena-se todos os *contigs* da montagem do maior para o menor. Em seguida, começando pelo maior *contig*, o tamanho dos *contigs* é somado até que se atinja

metade do tamanho total da montagem. O N50 desta montagem é definido como o tamanho do último *contig* deste conjunto [60]. Em geral, quanto maior o N50, melhor é a montagem. A qualidade também pode ser avaliada através do mapeamento das *reads*, isto é, através da sobreposição destas nos recém-montados *contigs*. Espera-se que os *contigs* apresentem uma boa quantidade de *reads* sobrepostas em toda sua extensão, ou seja, que todas suas regiões tenham cobertura alta.

### 1.3.5 Predição e anotação de ORFs

A predição de genes é feita com base nas ORFs (fase de leitura aberta, do inglês *open reading frame*): uma região genômica definida por um códon de início e um códon de parada, grande o bastante para codificar uma proteína. Embora nem todas as ORFs representem genes codificadores de proteína, elas funcionam como uma boa evidência da existência destes [61]. Existem uma série de softwares desenvolvidos para a predição de ORFs, entre eles o GLIMMER [62] e o *Genome Reverse Compiler* (GRC) [63].

Anotar um gene significa atribuir função biológica a ele. Na anotação de genes, infere-se que genes similares apresentam funções biológicas semelhantes, portanto suas anotações são compartilhadas. Para verificar a similaridade entre uma ORF e uma sequência já anotada, é necessário alinhá-las [61], ou seja, arranjá-las lado a lado de forma que as regiões similares entre elas estejam justapostas. Existem dois tipos de de alinhamento: o alinhamento global, em que sequências inteiras são comparadas; e o alinhamento local, em que fragmentos curtos das duas sequências são comparados. Uma representação gráfica dos dois tipos de alinhamento é mostrada na Figura 7.

Uma das ferramentas mais utilizadas no alinhamento de sequências é o *Basic Local Alignment Search Tool* (BLAST) [65], uma ferramenta de alinhamento local. Com o crescimento dos bancos de dados e o aprimoramento da anotação dos genes depositados nestes bancos, a anotação de genes tem ficado cada vez mais confiável.

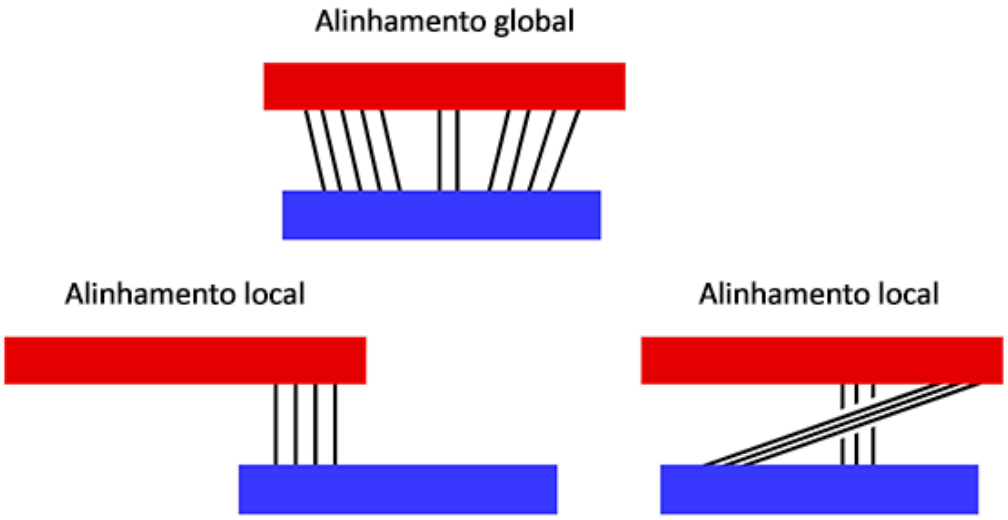



Figura 7: Representação esquemática de alinhamento global e local. Os retângulos de cores diferentes representam as sequências sendo alinhadas, e as linhas pretas mostram as regiões de similaridade entre as sequências. Fonte: adaptada de `http://www.pitt.edu/~mcs2/teaching/biocomp/tutorials/global.html`.

## 1.4 Ferramentas de bioinformática

A seguir, serão descritas as principais ferramentas utilizadas para a análise de bioinformática deste trabalho.

### 1.4.1 FastQC

O FastQC [66] é uma ferramenta desenvolvida pelo Babraham Institute, especializada em testes de qualidade para sequências obtidas de sequenciadores de próxima geração. Dentre os testes feitos pela ferramenta estão a análise de Phred das bases, a verificação de sequências muito representadas (que podem indicar adaptadores de sequenciamento não removidas) e a análise do conteúdo GC das *reads*.

### 1.4.2 MIRA Assembler

O MIRA assembler [58] é um montador de fragmentos de DNA especializado em projetos em que as *reads* apresentam muitas repetições. A ferramenta também é bastante utilizada na montagem de genomas, em particular de pequenos, como os bacterianos, a partir de sequências de sequenciadores de próxima geração. Com o MIRA, é possível fazer montagens *de novo* e *mapping* a partir de sequências de uma única tecnologia de

sequenciamento ou ainda a partir de uma mistura de sequências de plataformas diferentes (montagem híbrida).

O MIRA é baseado em uma variação da abordagem *overlap-layout-consensus* (OLC), onde cada *read* é representada em um grafo como um vértice, e a sobreposição entre duas *reads* é uma aresta entre os vértices apropriados (Figura 8).

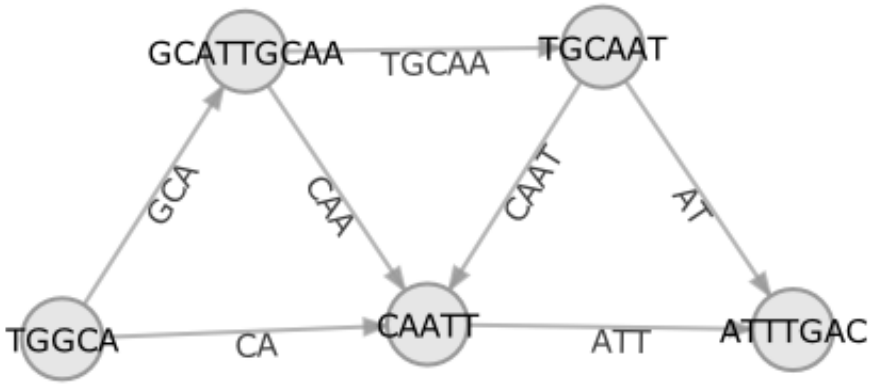



Figura 8: Representação esquemática de um grafo da abordagem *overlap-layout-consensus*. Cada vértice representa uma *read* inteira, e as arestas são as regiões de sobreposição entre duas *reads*. A sequência consenso é determinada pelo melhor caminho que percorre cada vértice uma única vez. Fonte: `http://gcat.davidson.edu/phast/olc.html`

### 1.4.3 Segemehl

O Segemehl [68] é uma ferramenta de mapeamento de *reads* baseada em árvores de sufixo desenvolvida na Universidade de Leipzig. Além de detectar *mismatches* (pareamentos de base errôneos), o segemehl também é capaz de identificar inserções e deleções.

### 1.4.4 Genome Reverse Compiler (GRC)

O GRC [63] é uma ferramenta que utiliza informações de organismos relacionados para a predição de ORFs e a anotação de genomas procarióticos. Para utilizar a ferramenta, o usuário precisa, previamente, criar um banco com os genomas evolutivamente próximos ao genoma estudado.

Em uma primeira etapa, o GRC procura por todas as possíveis ORFs dentro do genoma estudado. Em seguida, a ferramenta avalia a probabilidade de cada ORF representar um gene codificador de proteína. Alguns fatores considerados nessa avaliação são: se a composição de aminoácidos é característica de genes codificadores típicos do organismo, se a sequência é conservada em múltiplos organismos, se duas ou mais ORFs

se sobrepõem ou ainda se a ORF tem um tamanho característico de genes codificadores de proteína.

Após a determinação dos genes codifcadores de proteína, uma comparação com os genomas do banco é feita para realizar a anotação dos genes.

### 1.4.5 BLAST

O BLAST é uma ferramenta utilizada para avaliar a similaridade entre sequências de nucleotídeos ou aminoácidos e calcular a significância estatística desses pareamentos. Um dos principais usos do BLAST é inferir a relação evolutiva e funcional entre duas sequências distintas, sendo uma das principais ferramentas utilizadas na anotação de genes.

Para calcular a significância dos alinhamentos o BLAST utiliza um parâmetro chamado *expect value*, ou E-value. Esse parâmetro descreve o número esperado de alinhamentos que acontecem ao acaso ao se utilizar um banco de dados de um tamanho particular. Quanto mais próximo o E-value é de 0, maior a significância de um alinhamento, isto é, menor é a probabilidade dele ter ocorrido ao acaso.

### 1.4.6 Bancos de dados

Os bancos de dados biológicos são repositórios públicos de informações das ciências da vida, como sequências de nucleotídeos e aminoácidos, mapas metabólicos e dados filogenéticos. Os bancos de dados podem ou não ser curados, isto é, os dados submetidos por pesquisadores aos bancos podem ou não ser avaliados por curadores externos.

O conhecimento biológico é dividido entre diversos bancos especializados. Alguns dos bancos mais utilizados são: o GenBank [69], um repositório de genes do NCBI (*National center for biotechnology information*); o COG (*clusters of orthologous groups*) [70], um banco que atribui funções resumidas a conjuntos de genes ortólogos, também pertencente ao NCBI; o KEGG (*Kyoto encyclopedia of genomes and genes*) [71] um repositório de genomas e vias enzimáticas; o UniProtKB/Swiss-Prot [72], um banco curado de sequências de proteínas; e o VFDB (*Virulence factor database*) [73], um repositório especializado em sequências identificadas como fatores de virulência.

# 2 Objetivos

## 2.1 Justificativas

As micobactérias de crescimento rápido representam uma classe de micobactérias ubíquas resistentes a uma gama de antibióticos e desinfetantes. A separação desses organismos a nível de espécie costuma ser problemática, particularmente por apresentarem tanto características fenotípicas quanto sintomatologia de suas infecções virtualmente indistinguíveis.

Com o aumento dos surtos causados por essas bactérias, o desenvolvimento de novas técnicas de detecção e de predição de padrões de susceptibilidade é essencial para o controle eficiente de suas infecções. A análise genômica é uma ferramenta importante e que cada vez mais vem sendo utilizada na predição de alvos moleculares para a identificação e a inibição de bactérias patogênicas, porém os dados disponíveis em bancos de dados biológicos referentes às RGM ainda são escassos.

## 2.2 Objetivos

### 2.2.1 Objetivo Geral

- Realizar a análise do genoma da estirpe GO 06 de *Mycobacterium massiliense*, isolada durante o surto ocorrido entre os anos de 2005 e 2007 na região Centro-Oeste do país, de forma a se obter dados biológicos relevantes para o melhor entendimento do funcionamento e dos mecanismos de patogenicidade do organismo.

### 2.2.2 Objetivos Específicos

- Montar o genoma do isolado GO 06 de *Mycobacterium massiliense*.
- Predizer os genes codificadores de proteínas, tRNAs e rRNAs do isolado, e realizar

a classificação por ontologia de genes.

- Identificar os genes do isolado envolvidos na codificação de fatores de virulência.
- Traçar o mapa metabólico das vias responsáveis pela produção de sideróforos e de sistemas de secreção.
- Criar um banco de dados online de livre acesso com todos os dados biológicos gerados neste trabalho.

# 3 Material e Métodos

## 3.1 Sequenciamento do DNA e filtragem das *reads*

Durante o surto causado por RGM no estado de Goiás no período entre 2005 e 2007, isolados de infecções pós-cirurgicas foram recebidos pelo LACEN-GO. Dentre as estirpes, a estirpe nomeada "GO 06" foi escolhida para a análise genômica.

O genoma do isolado GO 06 foi sequenciado pela abordagem *whole genome shotgun* na plataforma de sequenciamento de próxima geração 454 GS-FLX Titanium da Roche. As *reads* foram extraídas dos dados brutos obtidos do sequenciador por meio de um *script* desenvolvido para esta finalidade (sff_extract) [74]. A filtragem das sequências foi feita pelo próprio *script* durante a extração, logo não houve necessidade da utilização de outras ferramentas de filtragem. A qualidade das *reads* extraídas foi avaliada pelo FastQC.

## 3.2 Montagem do genoma

Os *contigs* foram construídos por montagem *de novo* das *reads*. Para isso, foi utilizado o MIRA Assembler (versão 3.4.0) com os parâmetros padrão para montagens de sequências genômicas proveninentes da plataforma 454. Essa metodologia foi escolhida a partir da comparação das montagens feitas com a utilização de diversas metodologias: montagem *de novo* com o CAP3, AMOS e MIRA e montagem *mapping* com o MIRA.

Para verificar a integridade da montagem, o maior *contig* (aproximadamente 4.7 Mb), que provavelmente correspondia ao cromossomo do organismo, foi submetido a alinhamentos contra genomas de outras micobactérias. Esses alinhamentos de genoma contra genoma foram feitos pela ferramenta MUMmer (versão 3.0) [75], e com esses dados foi possível construir *dot plots* para cada uma das comparações. Os genomas escolhidos e seus respectivos números de acesso no NCBI são: *M. abscessus* ATCC 19977 (CU458896), *M. bovis* Pasteur 1173P2 (NC_008769), *M. leprae* Br4923 (NC_011896), *M. tuberculosis*

CDC1551 (NC_002755) e *Mycobacterium smegmatis* MC2 155 (NC_008596). Ademais, as *reads* foram mapeadas nos *contigs* obtidos por meio da ferramenta Segemehl, para análise da cobertura da montagem.

Em colaboração com o departamento de Ciência da Computação da Universidade de Brasília (CIC/UnB), foi desenvolvido um método para montagem dos plasmídeos a partir de sequências não utilizadas na montagem do cromossomo. Para verificar a similaridade dos plasmídeos montados com outros plasmídeos de micobactérias obtidos do NCBI, foram feitos alinhamentos pelo BLAST.

## 3.3 Predição de ORFs e anotação preliminar

A primeira etapa do *pipeline* de anotação foi a predição das ORFs, feita pelo GRC (versão 1.0), que também incluiu uma etapa de anotação preliminar. Todas as configurações de uso foram mantidas no padrão definido pelo programa. Os genomas utilizados na construção do banco do GRC e seus respectivos números de acesso no NCBI são: *M. abscessus* ATCC 19977, *M. bovis* AF2122/97 (NC_002945), *M. bovis* Mexico (NC_016804), *M. bovis* Pasteur 1173P2 (NC_008769), *M. bovis* Tokyo 172 (NC_012207), *M. leprae* Br4923, *M. leprae* TN (NC_002677), *M. tuberculosis* H37Rv (NC_000962), *M. tuberculosis* CDC1551, *M. tuberculosis* H37Ra (NC_009525) e *M. tuberculosis* CDC5180 (NC_017522).

Nesta etapa, os rRNAs e tRNAs também foram anotados. A identificação destes ncRNAs foi feita por meio de ferramentas online, RNAmmer (versão 1.2) [76] e tRNAscan-SE (versão 1.21) [77], respectivamente.

## 3.4 *Pipeline* de anotação e análise funcional

O *pipeline* desenvolvido para a anotação e análise funcional do genoma do isolado GO 06 é mostrado na Figura 9. O *pipeline* é dividido em quatro grandes etapas: (1) o aprimoramento das anotações das ORFs; (2) a identificação de fatores de virulência; (3) a análise de vias metabólicas de fatores de virulência de interesse; e (4) o mapeamento das vias metabólicas estudadas.

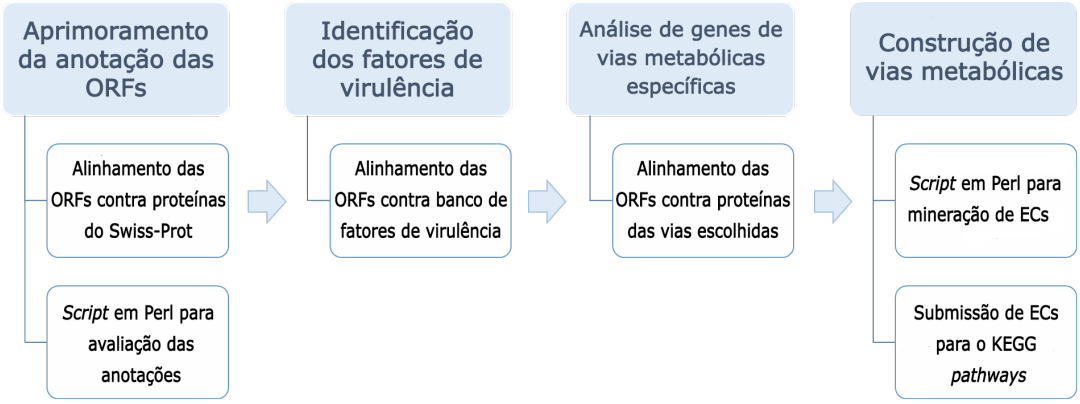



Figura 9: *Pipeline* para a anotação do genoma do isolado GO 06. Fonte: elaborada pelo autor.

### 3.4.1 Aprimoramento das anotações

Como o banco de dados utilizado pelo GRC para a inferência da anotação foi limitado, fez-se necessário aprimorar a anotação das ORFs identificadas, no intuito de se obter um conjunto de dados melhor caracterizado. As ORFs foram então alinhadas contra um banco composto por todas as sequências de proteínas bacterianas curadas do UniprotKB/Swiss-Prot (329.037 sequências, em abril de 2013). O alinhamento foi feito pelo BLAST, com um E-value limite de 1e-5.

Em seguida, um *script* em Perl foi construído para comparar as anotações feitas pelo GRC e pelo BLAST. A presença de palavras-chave como *hypothetical* ou *putative*, que indicam anotações genéricas, penalizavam a pontuação de uma anotação. Já a presença de informações complementares, como o nome do gene (GN) ou o *enzyme comission* (EC), aumentavam o a pontuação da anotação. Essa metodologia de pontuação foi desenvolvida para penalizar sequências com anotações pobres e favorecer anotações mais completas. O *script* também analisava cada possível alinhamento do BLAST, não se limitando apenas ao melhor alinhamento, de forma que cada anotação poderia ser escolhida dentre várias alternativas. Testes iniciais com este *script* sugeriram que penalizando palavras-chave em 2 pontos e adicionando meio ponto por cada informação complementar era possível evitar a maioria das anotações incompletas, portanto estes valores foram utilizados para a pontuação das anotações deste trabalho.

Ao final da etapa de pontuação, a anotação com o maior valor era mantida. Caso duas anotações apresentassem a mesma pontuação, as anotações provenientes do BLAST eram priorizadas. No caso de dois ou mais alinhamentos do BLAST com a mesma pontuação, o que apresentasse o menor E-value era escolhido.

### 3.4.2 Classificação dos genes em COGs

A classificação dos genes em COGs foi feita por meio de seu alinhamento, utilizando o BLAST, contra o banco de sequências do COG. A cada ORF que apresentasse pelo menos 70% de similaridade com alguma sequência do banco COG era atribuído o código desta sequência. Desta forma, foi possível criar um tabela que correlacionava cada ORF com um código do banco COG. Um *script* em Perl foi então desenvolvido para fazer o cruzamento dos dados da tabela criada com uma tabela disponível no COG que atribui a cada código uma classificação específica.

### 3.4.3 Identificação de fatores de virulência

Para determinar os fatores de virulência presentes no genoma do isolado GO 06, fez-se um alinhamento de todas as ORFs preditas contra sequências de proteínas obtidas no VFDB. O alinhamento foi feito pelo BLAST, com um E-value limite de 1e-5. Todas as ORFs que apresentaram pelo menos 70% de similaridade com alguma sequência proveniente do VFDB foram classificadas como fatores de virulência, indedependentemente da sua anotação.

### 3.4.4 Mapeamento de vias metabólicas

Informações sobre as vias metabólicas envolvidas na produção de sideróforos e das proteínas dos sistemas de secreção foram obtidas do KEGG. As sequências de todos os genes de famílias gênicas associadas a estas vias foram obtidas e utilizadas na construção de um banco de dados do BLAST. Em seguida, por meio de alinhamentos feitos pelo BLAST, foram identificadas as sequências que possuiam pelo menos 70% de similaridade com alguma sequência do banco de dados. Dessa forma, foi possível delinear os elementos de cada via metabólica presentes no genoma do isolado GO 06.

O código EC dos genes identificados foi obtido através de um *script* desenvolvido para manipulação de *flatfiles*. Esses códigos foram submetidos ao KEGG *pathways*, uma ferramenta do banco KEGG que lida com vias metabólicas, para a construção das vias específicas de *M. massiliense* GO 06. Não foi possível construir um mapa metabólico para a via de produção do sistema de secreção do tipo 7, já que este ainda é pouco caracterizado e não existem informações sobre este sistema no KEGG.

## 3.5 Banco de dados e disponibilização no domínio web

Com os dados de montagem e anotação do genoma de *M. massiliense* GO 06 foram criados bancos de dados em MySQL. O acesso livre a estes bancos foi disponibilizado através de uma página criada em HTML, PHP e Javascript e hospedada no domínio do laboratório de Biologia Molecular da Universidade de Brasília. Além do banco de dados indexado, foram disponibilizadas para download no site todas as sequências das montagens e das ORFs preditas, bem como algumas das figuras presentes neste estudo.

# 4 *Resultados*

## 4.1 Sequenciamento do DNA e filtragem das *reads*

Um total de 584.619 *reads* foram obtidas por meio do sequenciamento 454. A análise da qualidade das sequências demonstrou a ausência de sequências muito representadas, que podem indicar a presença de adaptadores de sequenciamento, e qualidade média das bases (escala Phred) acima de 20 (Figura 10).

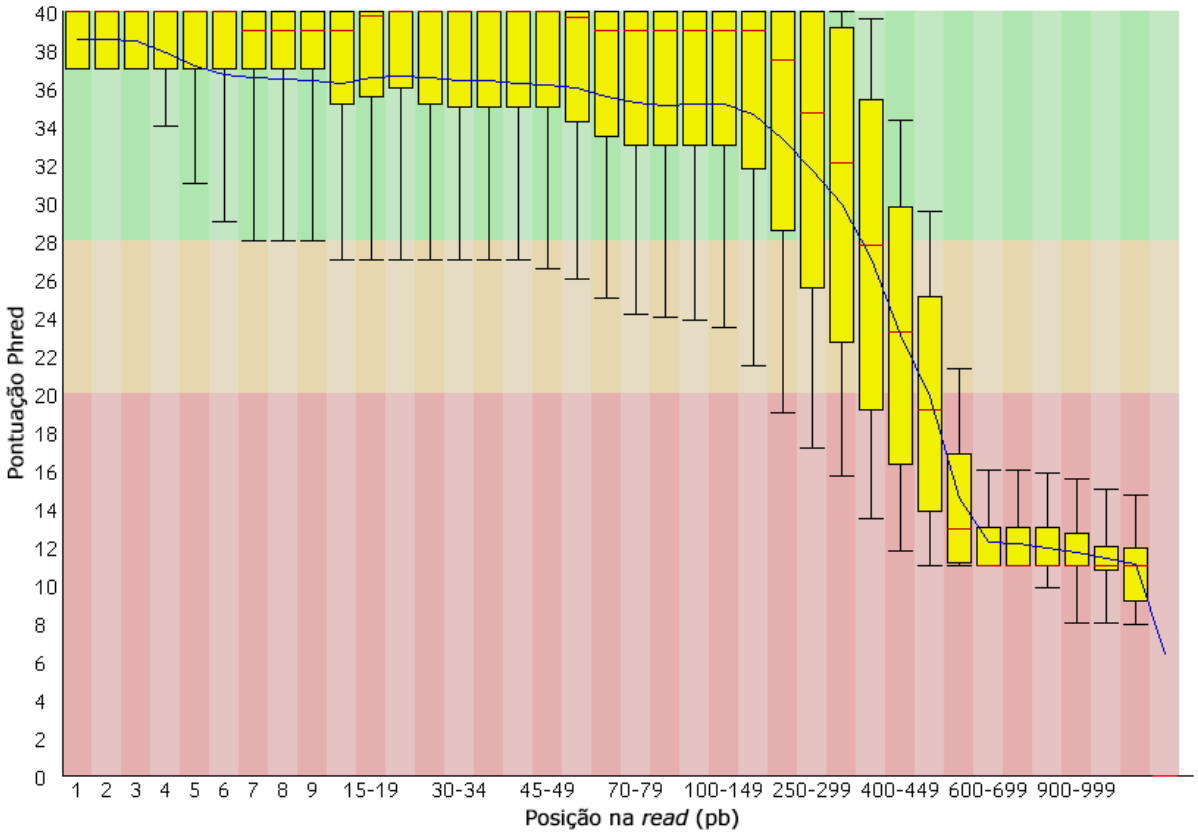



Figura 10: Pontuação Phred de cada base dos fragmentos sequenciados, ao longo de suas extensões. A linha central vermelha representa a mediana da qualidade, enquanto as caixas amarelas representam a amplitude inter-quartil da qualidade das sequências. A linha azul representa a média da qualidade. Fonte: adaptada da saída do programa FastQC.

## 4.2 Montagem do genoma

Os dados utilizados na comparação entre as ferramentas empregadas na montagem do genoma de *M. massiliense* GO 06 se encontram na Tabela 4. Na montagem escolhida (MIRA *de novo*), 576.506 *reads*, representando 98,6% do total sequenciado, foram montadas em 94 *contigs*. O maior *contig*, de 4.687.873 pb, foi identificado como o provável cromossomo deste isolado. Este cromossomo foi disponibilizado no repositório de genomas do NCBI com o número de acesso CP003699. Com os *contigs* restantes, foi possível montar dois grandes *contigs* circulares de aproximadamente 60 e 96 kb que provavelmente correspondem a plasmídeos. Estes foram nomeados "Plasmídeo I"e "Plasmídeo II", respectivamente.

Tabela 4: Comparação entre as ferramentas empregadas na montagem do cromossomo de *M. massiliense* GO 06

| | CAP3 | AMOS | MIRA (*de novo*) | MIRA (*mapping*) |
|---|---|---|---|---|
| Número de *contigs* | 343 | 97 | 94 | 1 |
| Tamanho médio dos *contigs* (pb) | 15.845 | 49.190 | 52.237 | 5.068.807 |
| Tamanho do maior *contig* (pb) | 814.179 | 443.471 | 4.687.873 | 5.068.807 |
| Número de *singlets* | 10.275 | - | 8.133 | 138.118 |

O mapeamento das *reads* no cromossomo e nos plasmídeos demonstrou boa cobertura em toda a extensão das montagens (Figura 11). O cromossomo apresentou uma média de 45x de cobertura em sua extensão. Já os plasmídeos I e II apresentaram uma média de aproximadamente 338x e 70x, respectivamente. A cobertura média e outras informações referentes à montagem do genoma de *M. massiliense* GO 06 se encontram resumidas na Tabela 5. A representação gráfica do cromossomo do isolado GO 06, bem como de seus dois plasmídeos, pode ser visualizada na Figura 12. É possível observar, no mapa do cromossomo, regiões de grande similaridade entre os genomas comparados, em particular a grande similaridade de *M. massiliense* com *M. abscessus*. É possivel observar também que enquanto o GC *skew* do cromossomo apresenta distribuição normal, o *skew* dos plasmídeos é bastante variado ao longo de sua extensão.

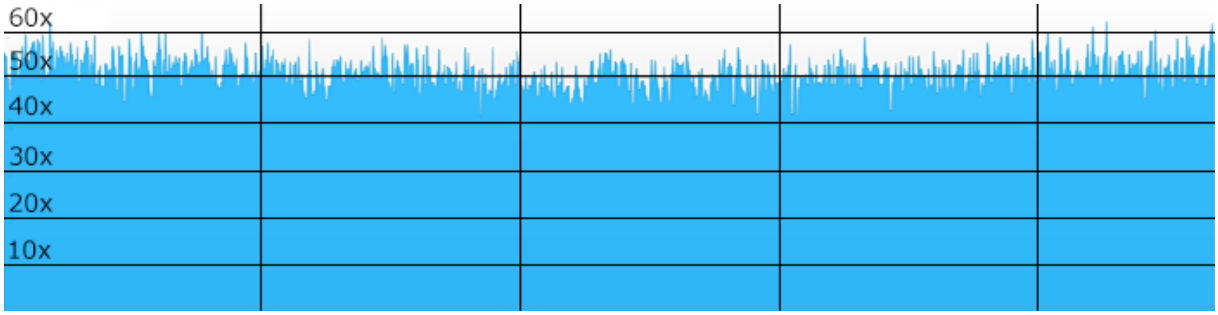
As análises *dot plot* (Figura 13) demonstraram grande similaridade entre o cromossomo de *M. massiliense* GO 06 e *M. abscessus* ATCC 19977 (Figura 13a), o que, juntamente com alto N50, sugere uma boa qualidade da montagem. Além disso, estas análises

mostram os variados níveis de similaridade do cromossomo montado com cromossomos de outras micobactérias, particularmente a maior similaridade de *M. massiliense* com bactérias saprofíticas (*M. smegmatis*) quando comparada a similaridade com bactérias patogênicas do complexo *M. tuberculosis* (*M. tuberculosis* e *M. bovis*).

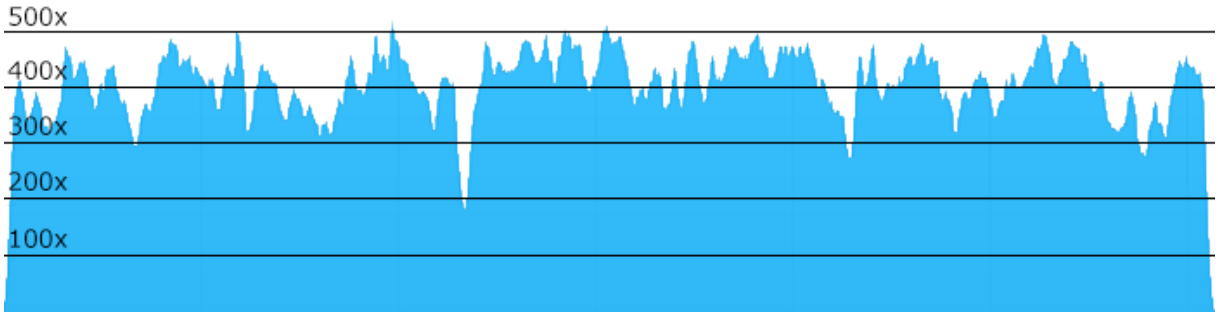
Tabela 5: Características gerais da montagem do genoma de *M. massiliense* GO 06

| | Cromossomo | Plasmídeo I | Plasmídeo II |
|---|---|---|---|
| Tamanho (pb) | 4.687.873 | 61.416 | 96.251 |
| Número de reads utilizadas na montagem | 501.303 | 51.156 | 22.022 |
| N50 | 4.687.873 | 61.120 | 5.210 |
| Cobertura média (número de reads) | 45,08 | 338,66 | 69,52 |
| Conteúdo GC (%) | 64,3 | 62,7 | 63,8 |

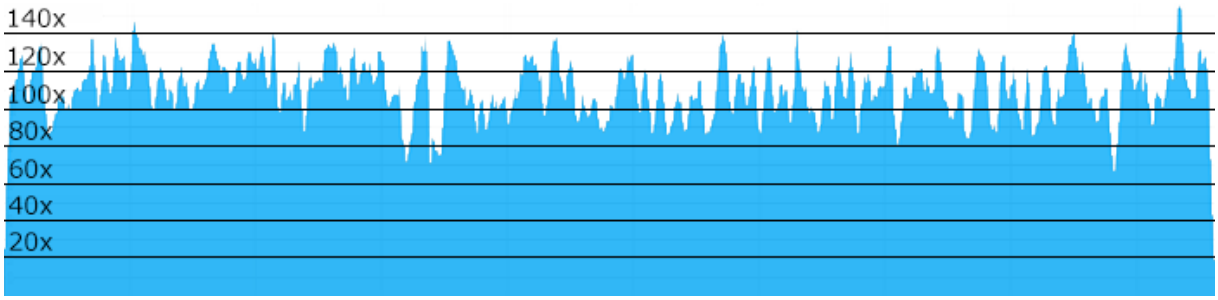
As análises *in silico* com os plasmídeos demonstraram que o plasmídeo I é altamente similar (99%) ao plasmídeo pMAB01 de *M. abscessus* subsp. *bolletii* (CP003376). O plasmídeo II não se mostrou similar a nenhum plasmídeo conhecido de micobactérias.




(a) Cromossomo

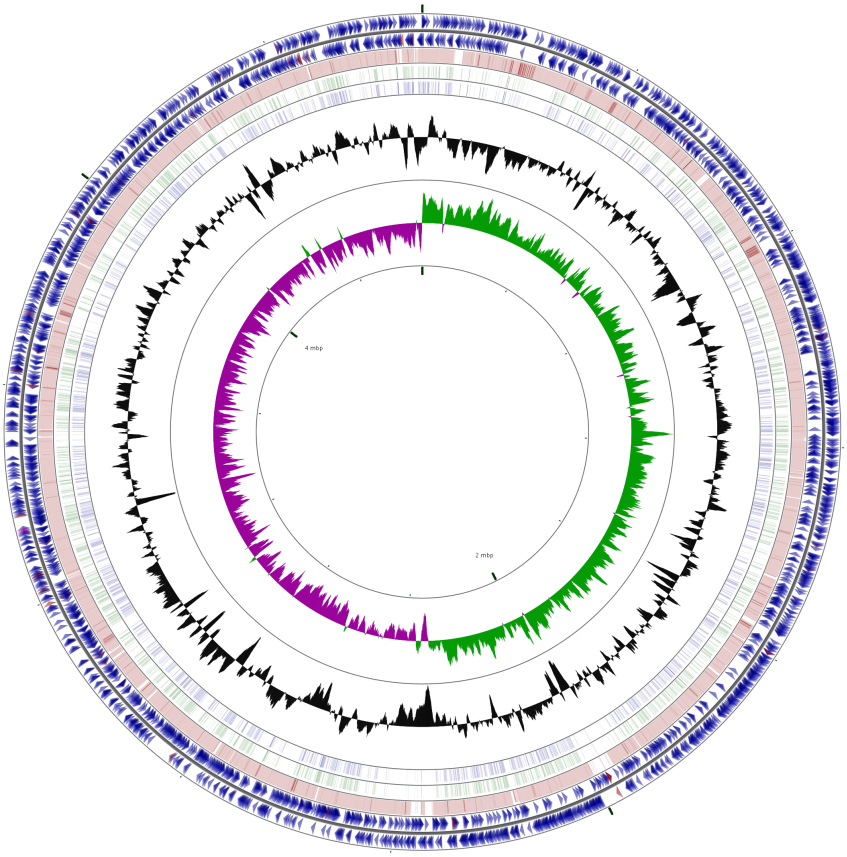



(b) Plasmídeo I

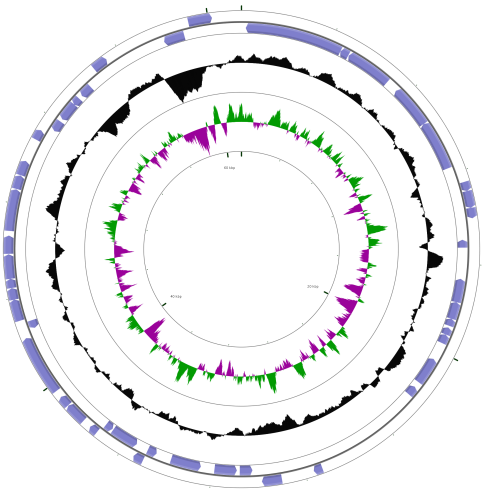



(c) Plasmídeo II

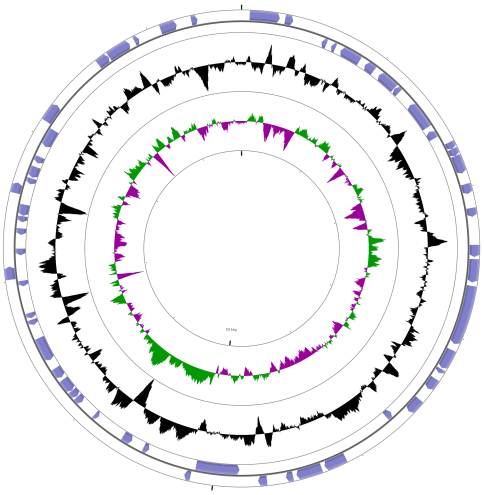
Figura 11: Gráfico de cobertura do cromossomo e plasmídeos de *M. massiliense* GO 06. O eixo x representa a sequência das montagens, ao longo de sua extensão, enquanto o eixo y representa a cobertura de uma determinada região. Figura adaptada da saída do programa *Savant Genome Browser*



(a) Cromossomo

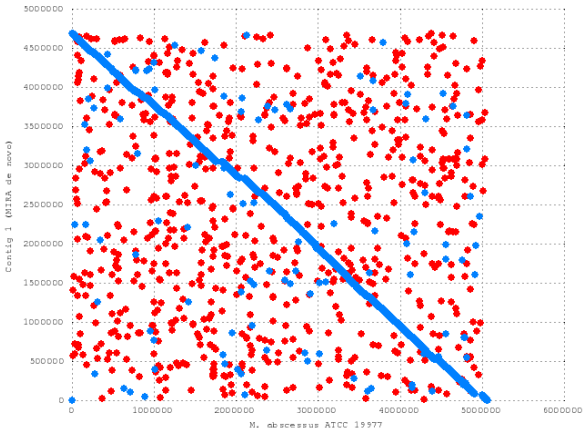


(b) Plasmídeo I

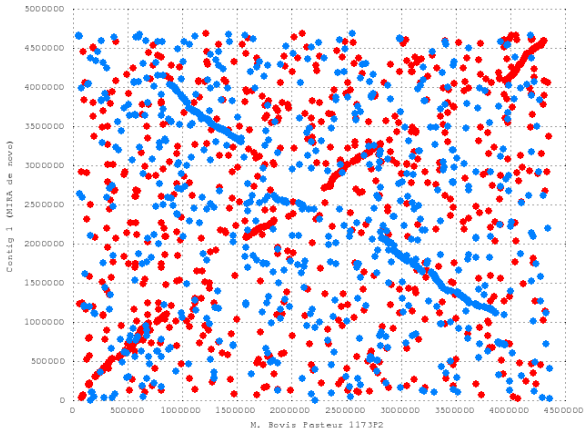


(c) Plasmídeo II

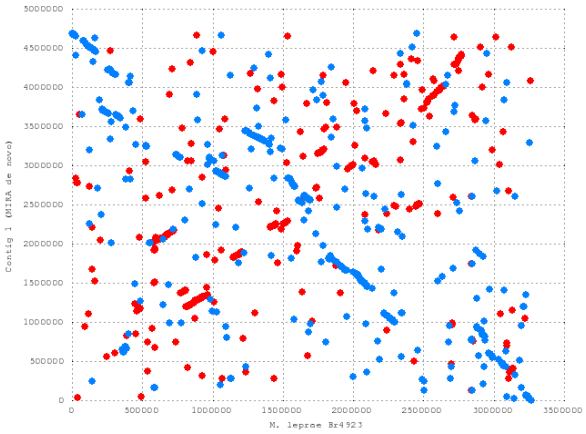
Figura 12: Visualização do genoma de *M. massiliense* GO 06. Cada círculo, de dentro para fora, representa: o cromossomo círcular ou o plasmídeo, o GC *skew* (*skew plus* em verde e *skew minus* em rosa), conteúdo de GC e as ORFs identificadas. No cromossomo, existem ainda três círculos, posicionados entre as ORFs identificadas e o conteúdo de GC, que representam a análise de similaridade do mesmo com cromossomos de *M. bovis* Pasteur 1173P2 (azul), *M. tuberculosis* CDC 1551 (verde) e *M. abscessus* ATCC 19977 (vermelho). A intensidade da cor destes círculos é diretamente proporcional à pontuação de similaridade do BLAST.



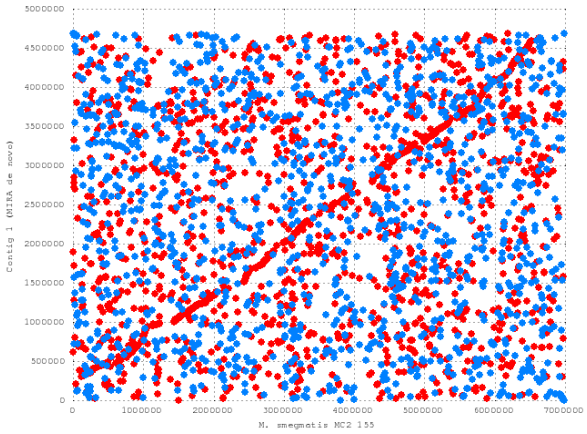

(a) *Mycobacterium abscessus* ATCC 19977



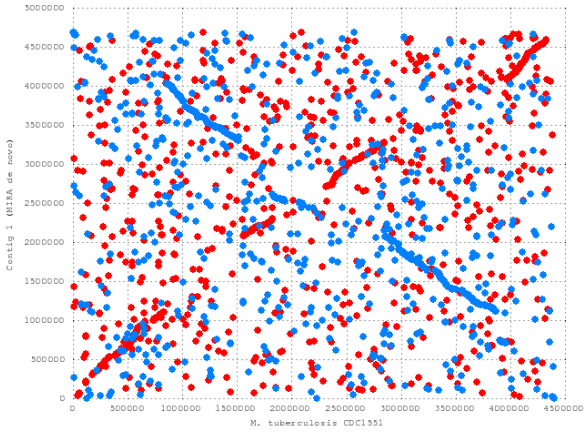

(b) *Mycobacterium bovis* Pasteur 1173P2




(c) *Mycobacterium leprae* Br4923




(d) *Mycobacterium smegmatis* MC2 155




(e) *Mycobacterium tuberculosis* CDC 1551

Figura 13: Análise *dot plot* entre o cromossomo de *M. massiliense* GO 06 e outras micobactérias. Os pontos em vermelho representam o alinhamento na direção senso, enquanto os azuis representam o alinhamento na direção antisenso. Em todas as figuras, o eixo y corresponde ao cromossomo montado do isolado GO 06, enquanto o eixo x corresponde ao genoma do organismo identificado pela legenda.

## 4.3 Anotação

### 4.3.1 Predição e anotação de ORFs

No total, foram preditas 4.131 ORFs no genoma de *M. massiliense* GO 06. No cromossomo, 3.386 (84,2%) ORFs conservadas foram identificadas, sendo possível atribuir função biológica a 2.389 (70,6%) delas. Já nos plasmídeos, 47 (97,9%) e 58 (96,7%) ORFs puderam ser anotadas, respectivamente. A Tabela 6 resume as informações referentes à anotação do genoma, destacando a diferença na quantidade de ORFs anotadas antes e depois da aplicação do *pipeline* de aprimoramento. É possível notar que houve um aumento na quantidade de ORFs anotadas após este procedimento, principalmente nos plasmídeos. Além das ORFs, foi possível anotar 46 tRNAs e um operon de rRNA.

Tabela 6: Número de ORFs identificadas no genoma de *M. massiliense* GO 06 e informações sobre sua anotação antes e depois da aplicação do *pipeline* de aprimoramento.

| | Cromossomo | | Plasmídeo I | | Plasmídeo II | |
|---|---|---|---|---|---|---|
| | Antes | Depois | Antes | Depois | Antes | Depois |
| Número total de ORFs | 4.023 | | 48 | | 60 | |
| ORFs conservadas com função atribuída | 1.394 | 2.389 | 9 | 47 | 8 | 58 |
| ORFs conservadas de função genérica | 1.659 | 997 | 5 | 0 | 7 | 0 |
| ORFs não conservadas | 970 | 637 | 34 | 1 | 45 | 2 |

### 4.3.2 Classificação dos genes em COGs

Das 4.023 ORFs identificadas no cromossomo de *M. massiliense* GO 06, 3.215 puderam ser classificadas de acordo com as categorias do COG, correspondendo a 94,9% de todas as ORFs conservadas (Figura 14a). A maior parte destas ORFs está envolvida no metabolismo de aminoácidos (9,2%), na transcrição (8,8%) e no metabolismo de metabólitos secundários (8%).

Os plasmídeos também foram classificados de acordo com o COG. No plasmídeo I, 33 ORFs (70,2% das ORFs conservadas) foram distribuídas em 13 categorias (Figura 14b). 14 ORFs do plasmídeo II (24,1% das ORFs conservadas) foram dividas em apenas 5 categorias (Figura 14c). No plasmídeo I, a maior parte dos genes que puderam ser classificados estavam relacionados ao controle do ciclo celular, e aos processos de recombinação, reparo

e transcrição. Já no plasmídeo II, 60% dos genes estavam relacionados ao transporte intracelular e secreção e à replicação, recombinação e reparo.

### 4.3.3 Identificação de fatores de virulência e mapeamento das vias metabólicas

De todas as ORFs anotadas, 826 puderam ser classificadas como fatores de virulência. Destas, 124 (15%) foram correlacionadas ao metabolismo de lipídios, 191 (23,1%) à via de produção de sideróforos e 114 (13,8%) à via de produção de sistemas de secreção bacterianos. Além desses sistemas, foi observado um grande número de genes envolvidos na biossíntese da parede celular, como os responsáveis pela síntese dos ácidos micólicos, e de resistência a antibióticos. A distribuição dos genes de virulência nas categorias do COG está representada na Tabela 7.

A maioria dos genes relacionados à via de biossíntese dos sideróforos pertenciam a família *entA* (81 cópias), um precursor de sideróforos como a mixoquelina e a enteroquelina. A Figura 15 mostra a distribuição destes genes nas famílias gênicas que formam a via metabólica. O mapa desta via pode ser visualizado na Figura 16, onde as famílias gênicas presentes no genoma de *M. massiliense* GO 06 são marcadas em verde. É possível notar que a maioria das famílias gênicas que formam esta via em particular estão presentes no mapa metabólico do isolado GO 06.

Além de genes relacionados aos sistemas Sec-SRP e Tat, genes de quatro tipos de sistema de secreção foram identificados no genoma de *M. massiliense* (I, III, IV e VI), grande parte deles relacionados ao T1SS (49 genes) e ao T3SS (27 genes). Apesar de 49,5% dos genes identificados estarem relacionados ao T1SS, todos os 49 genes caracterizados como tal representavam cópias do gene *hlyB*, responsável pela codificação de um transportador molecular (transportador ABC). A distribuição dos genes do isolado GO 06 relacionados aos sistemas de secreção nas famílias gênicas que compõem esta via pode ser visualizada na Figura 17. A Figura 18 representa o mapa da via metabólica de produção dos sistemas de secreção bacterianos, onde as famílias gênicas presentes no genoma de *M. massiliense* GO 06 foram marcadas em verde. Foi observado que, com excessão do sistema Sec-SRP, poucas famílias gênicas eram representadas no mapa metabólico do isolado GO 06 para as vias dos sistemas de secreção bacterianos.

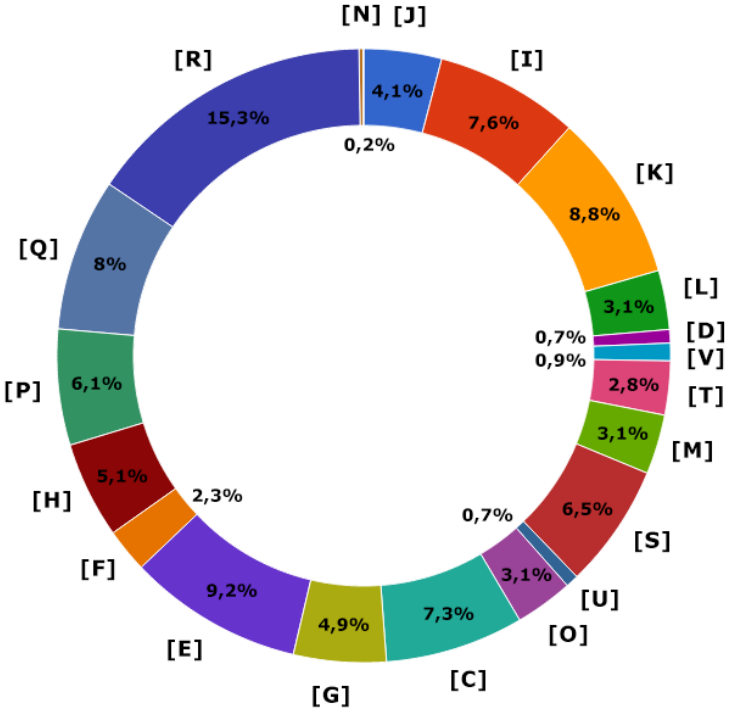
Tabela 7: Distribuição dos genes de virulência de *M. massiliense* GO 06 nas categorias do COG. Para melhor caracterizar o perfil de patogenicidade do organismo, três novas categorias foram criadas: "Patogênese", "Resposta ao estresse" e "Biossíntese de antibióticos".

| Categoria | Cromossomo | Plasmídeo I | Plasmídeo II |
|---|---|---|---|
| Tradução, estrutura e biogênese de ribossomos | 5 | 0 | 0 |
| Transcrição | 58 | 0 | 0 |
| Replicação, recombinação e reparo | 5 | 8 | 0 |
| Controle do ciclo celular, divisão celular e partição de cromossomo | 21 | 0 | 1 |
| Patogênese | 79 | 1 | 6 |
| Mecanismos de transdução de sinal | 22 | 0 | 0 |
| Biogênese da parede ou envelope celular | 9 | 0 | 0 |
| Resposta ao estresse | 55 | 1 | 1 |
| Modificações pós-traducionais, chaperonas e *turn-over* | 27 | 0 | 0 |
| Produção e conversão de energia | 14 | 0 | 0 |
| Transporte e metabolismo de carboidratos | 37 | 0 | 0 |
| Transporte e metabolismo de aminoácidos | 49 | 0 | 0 |
| Transporte e metabolismo de nucleotídeos | 7 | 0 | 0 |
| Transporte e metabolismo de coenzimas | 19 | 0 | 0 |
| Transporte e metabolismo de lipídios | 124 | 0 | 0 |
| Transporte e metabolismo de íons inorgânicos | 46 | 0 | 1 |
| Transporte e metabolismo de metabólitos secundários | 66 | 0 | 0 |
| Biossíntese de antibióticos | 25 | 0 | 0 |
| Apenas função geral predita | 139 | 0 | 0 |

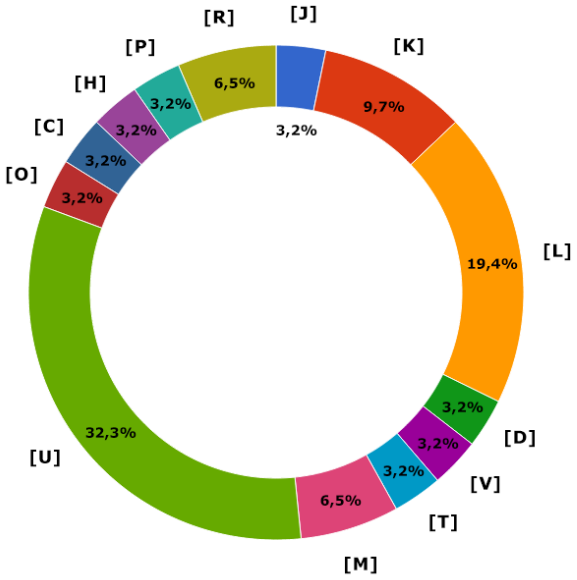
Além disso, foi possível identificar 15 genes relacionados ao T7SS, como mostra a Figura 19. Um esquema mostrando a organização dos genes relacionados ao T7SS no cromossomo do isolado GO 06 pode ser visualizado na Figura 20. A maior parte dos genes identificados relacionados a este sistema foram encontrados em sequência e distribuídos em dois *loci* distintos. Com exceção de um gene do segundo *locus*, todos os outros genes identificados apresentavam homologia ao ESX-3 ou ao ESX-4 de *M. tuberculosis* (Tabela 8).

Tabela 8: Principais genes do T7SS identificados em *M. massiliense* GO 06 e seus respectivos homólogos no genoma de *M. tuberculosis*.

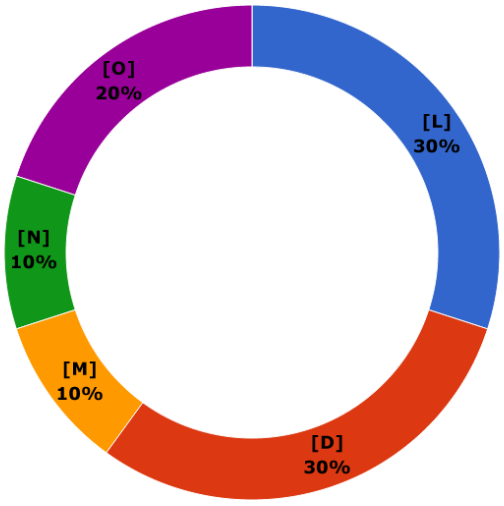
| *M. massiliense* | *locus* | *M. tuberculosis* | T7SS |
|---|---|---|---|
| EccB | I | Rv0283 | ESX-3 |
| EccC | I | Rv0284 | ESX-3 |
| EccD | I | Rv0290 | ESX-3 |
| EccE | I | Rv0292 | ESX-3 |
| EccA | II | Rv0282 | ESX-4 |
| EccB | II | Rv3450c | ESX-4 |
| EccC | II | Rv3447c | ESX-4 |
| EccD | II | Rv3448 | ESX-4 |
| EccE | II | Rv1797 | ESX-5 |




(a) Cromossomo




(b) Plasmídeo I




(c) Plasmídeo II

Figura 14: Distribuição das ORFs anotadas do genoma de *M. massiliense* GO 06 em categorias COG. Categorias sem representação foram omitidas. Nas legendas, cada letra entre colchetes representa uma categoria COG, que é acompanhada da respectiva porcentagem de ORFs classificadas como tal. [A] Modificação e processamento de RNA; [C] Produção e conversão de energia; [D] Controle do ciclo celular, divisão celular e partição de cromossomo; [E] Transporte e metabolismo de aminoácidos; [F] Transporte e metabolismo de nucleotídeos; [G] Transporte e metabolismo de carboidratos; [H] Transporte e metabolismo de coenzimas; [I] Transporte e metabolismo de lipídios; [J] Tradução, estrutura e biogênese de ribossomos; [K] Transcrição; [L] Replicação, recombinação e reparo; [M] Biogênese da parede ou envelope celular; [N] Motilidade celular; [O] Modificações pós-traducionais, chaperonas e *turn-over*; [P] Transporte e metabolismo de íons inorgânicos; [Q] Transporte e metabolismo de metabólitos secundários; [R] Apenas função geral predita; [S] Função desconhecida; [T] Mecanismos de transdução de sinal; [U] Transporte intracelular, secreção e vesículas; [V] Mecanismos de defesa.

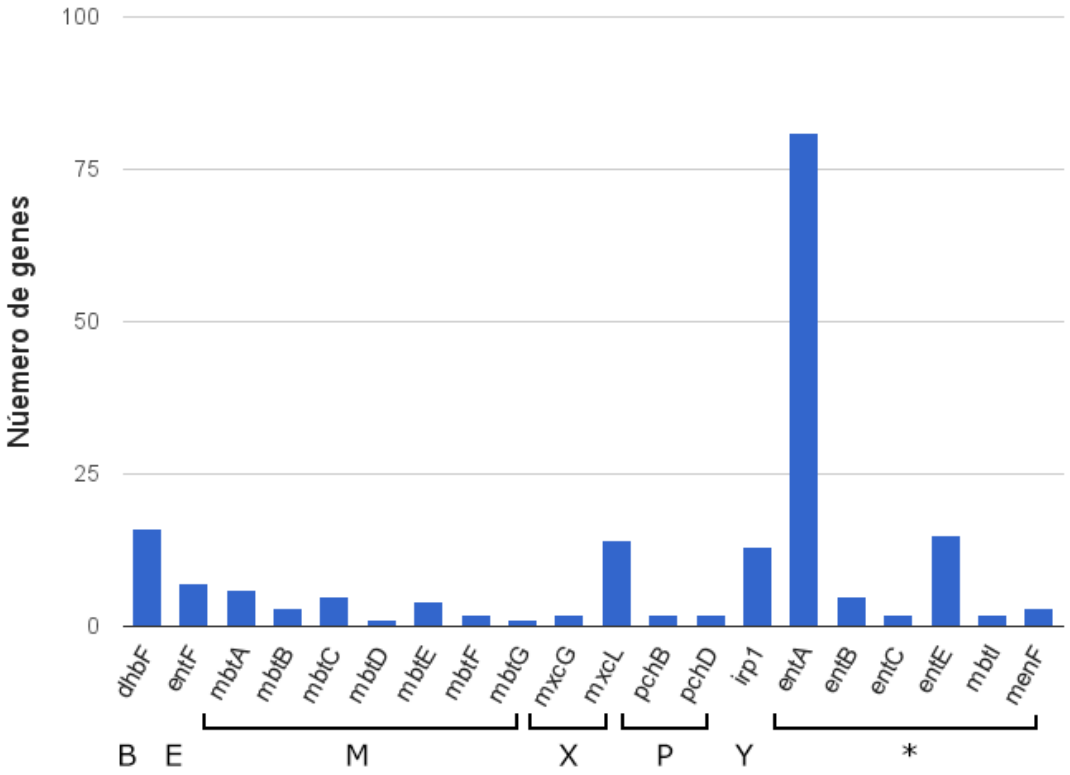



Figura 15: Distribuição dos genes de *M. massiliense* relacionados à produção de sideróforos. O eixo y representa o número de genes encontrados no genoma pertencentes a uma determinada família gênica. Um código foi atribuído a cada família gênica, representando o sideróforo em cuja produção estão envolvidas: Bacilibactina (B), Enteroquelina (E), Micobactina (M), Mixoquelina (X), Pioquelina (P) e Yersiniabactina (Y). As famílias gênicas marcadas com "*" ou estão envolvidas na produção de precursores de sideróforos, ou na produção de mais de uma das moléculas previamente citadas.

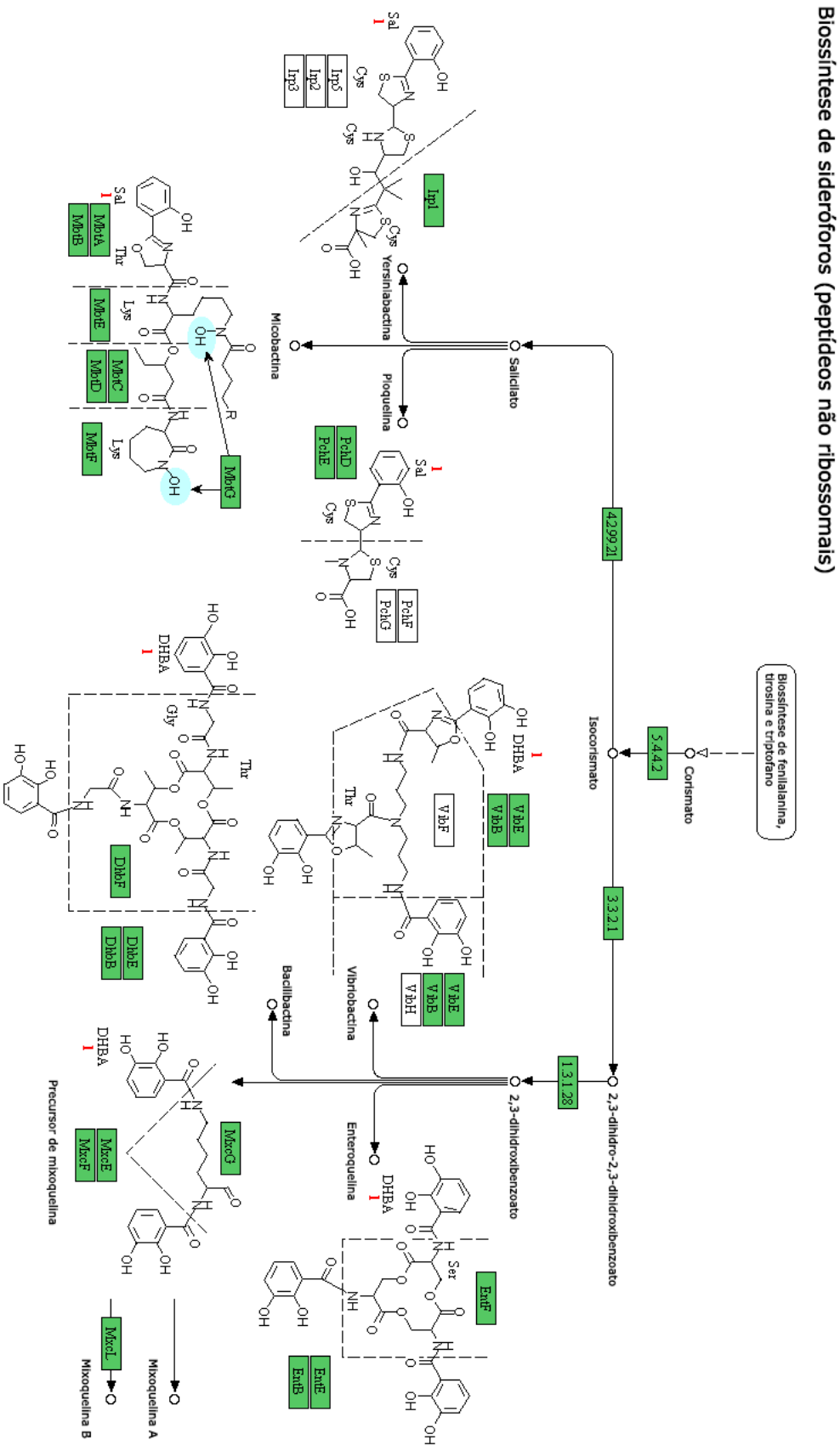



Figura 16: Mapa metabólico da produção de sideróforos. As famílias de genes com representatividade no genoma de *M. massiliense* GO 06 estão marcados em verde.

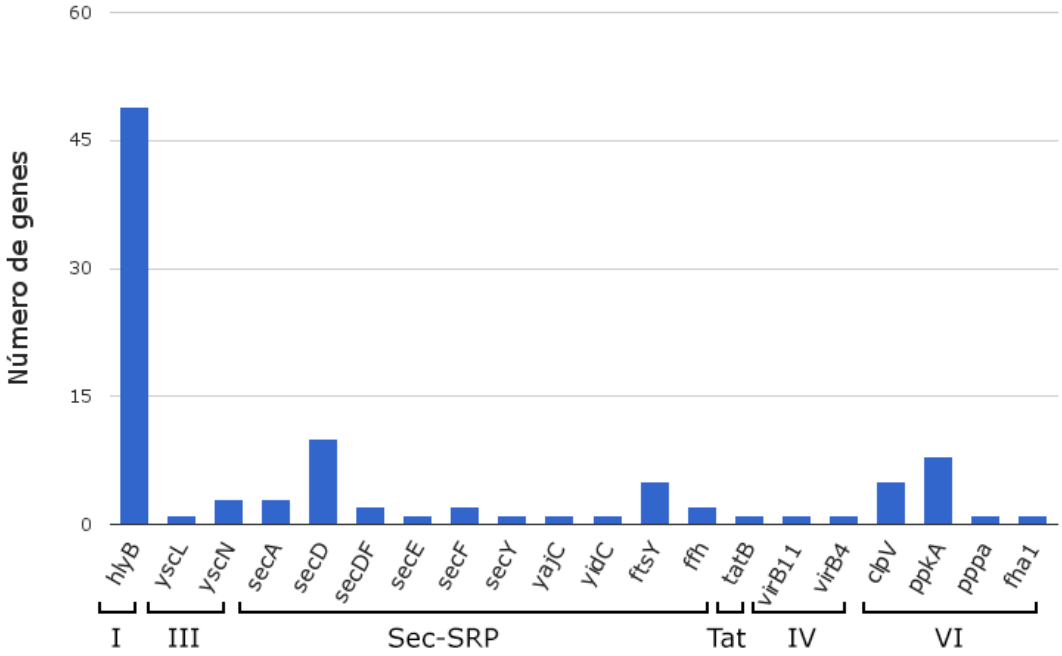



Figura 17: Distribuição dos genes de *M. massiliense* relacionados à produção de proteínas dos sistemas de secreção bacterianos. O eixo y representa o número de genes encontrados no genoma pertencentes a uma determinada família gênica. As legendas indicam o sistema de secreção em cuja produção as famílias gênicas estão envolvidas.

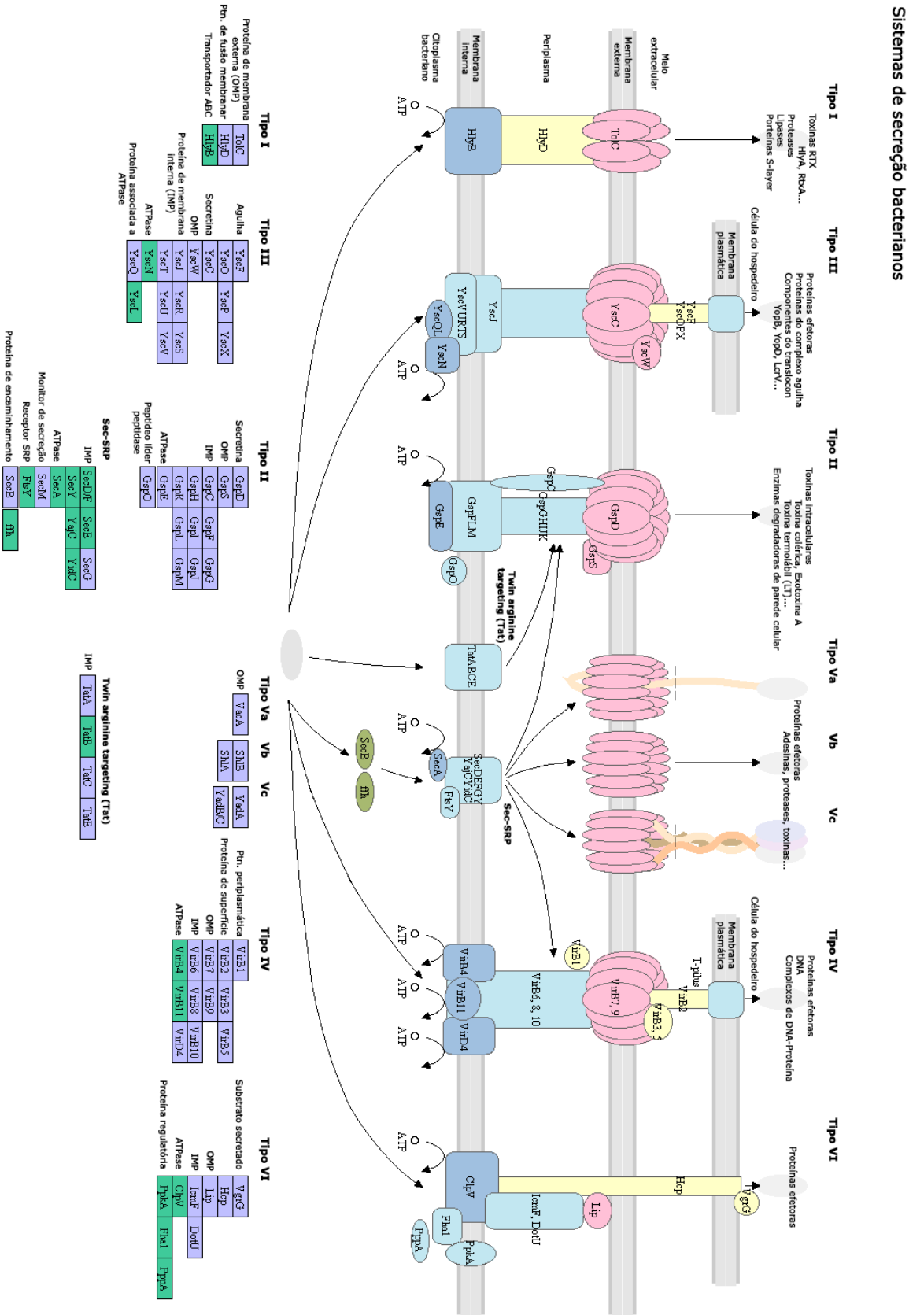



Figura 18: Mapa metabólico da produção de proteínas dos sistemas de secreção bacterianos. As famílias de genes com representatividade no genoma de *M. massiliense* GO 06 estão marcados em verde.

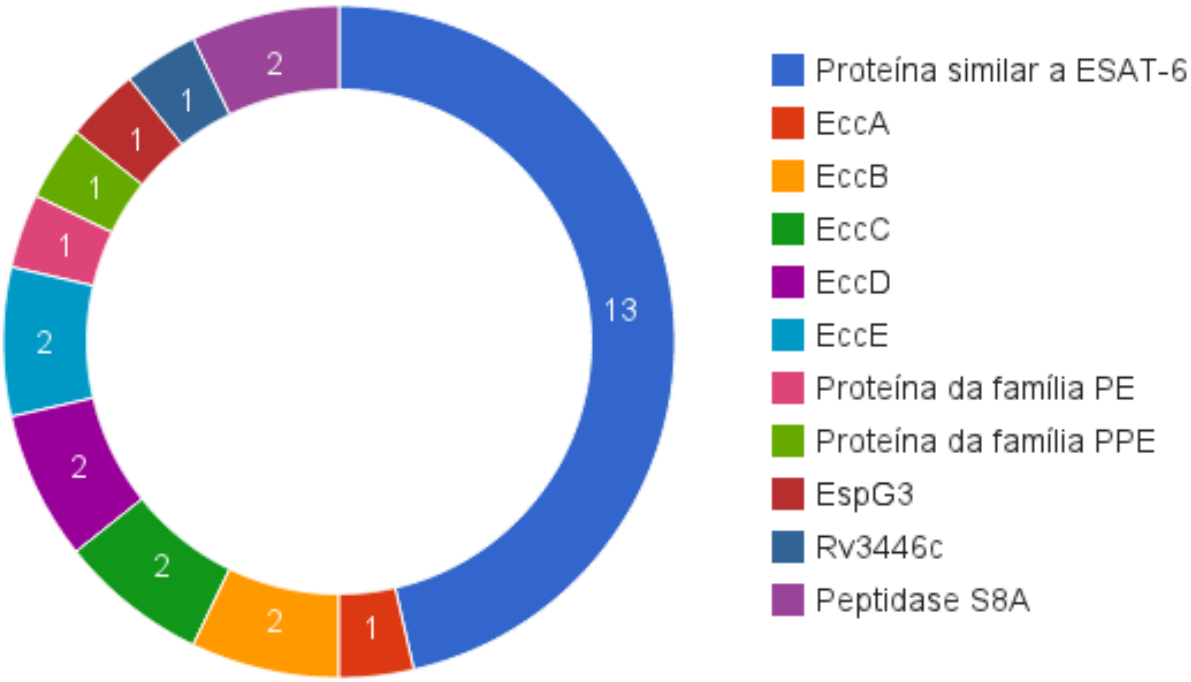



Figura 19: Distribuição dos genes de *M. massiliense* relacionados à produção de proteínas do T7SS.

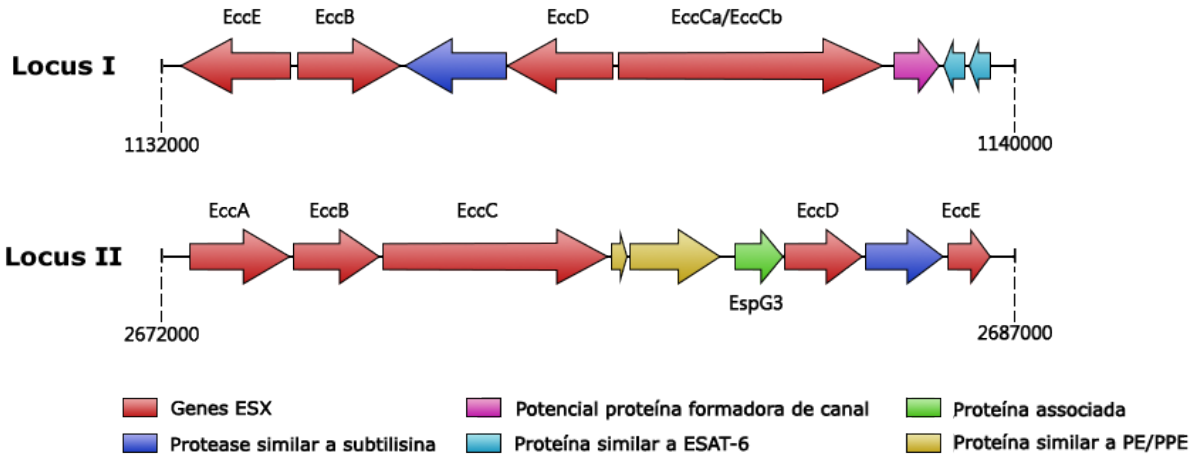



Figura 20: Organização dos genes relacionados ao T7SS no cromossomo de *M. massiliense* GO 06.

## 4.4 Divulgação científica

Durante a condução deste os resultados obtidos foram divulgados em três publicações: nos anos de 2012 e 2013 houve a apresentação e publicação de partes do projeto no *Brazilian Symposium on Bioinformatics* [78, 79], congresso de abrangência internacional referência na área de bioinformática. Com a deposição do genoma no NCBI, uma publicação de anúncio também foi feita [80]. Além disso, o trabalho apresentado no Simpósio de Biologia Molecular realizado na Universidade de Brasília em 2012 foi apreciado com o título melhor painel.

Por fim, os dados referentes a este projeto foram disponibilizados em uma página de livre acesso hospedada no servidor da Universidade de Brasília (`http://www.biomol.unb.br/massiliense/`). Além de informações gerais sobre o genoma de *M. massiliense* GO 06, a página também dá acesso a um banco de dados indexado, onde pode-se acessar e pesquisar as informações sobre a posição genômica, anotação e sequência de nucleotídeos e aminoácidos de cada ORF e ncRNA identificado. As sequências de nucleotídeos do cromossomo e plasmídeos montados durante o projeto estão disponíveis para *download* na página.

# 5 Discussão

Na última década, com o avanço dos métodos de cultura e identificação de bactérias, o número de casos clínicos reportados relacionados às micobactérias não tuberculosas têm aumentado. *M. abscessus*, uma micobactéria representativa do grupo das micobatérias de crescimento rápido, emergiu como um dos principais patógenos oportunistas, sendo considerado o agente etiológico de aproximadamente 65 a 80% de todos os casos reportados atribuídos às NTM [31]. No Brasil, três grandes surtos associados a micobactérias do complexo *M. abscessus* foram reportados em regiões distintas do país (Norte, Sudeste e Centro-Oeste) entre 2004 e 2009.

A identificação das micobactérias do complexo *M. abscessus* (*M. abscessus*, *M. massiliense* e *M. bolletii*) em nível de espécie é essencial para o tratamento adequado de suas infecções [31]. Apesar disso, a discriminação destas bactérias se mostra problemática, tanto pela comparação do perfil bioquímico das espécies que compõem este grupo, que pode variar de acordo com a estirpe estudada, quanto pela análise genotípica de genes individuais, que é insuficiente para a diferenciação destas espécies [81]. Ademais, como os sintomas das infecções causadas pelas micobactérias desse complexo são muito similares, estes não podem ser utilizados como critério para a determinação das espécies.

Um dos principais ensaios laboratoriais empregados na identificação de bactérias com perfis bioquímicos ambíguos é a análise genotípica a partir do sequenciamento do RNA ribossomal 16S [82]. Como as sequências de rRNA 16S das bactérias pertencentes ao complexo *M. abscessus* são idênticas, é possivel que casos clínicos relacionados a este grupo sejam erroneamente atribuídos a *M. abscessus*, seu representante mais conhecido. Dessa forma, além da impossibilidade de diferenciação pela análise genotípica do rRNA 16S, existe ainda a possibilidade de o número de infecções causadas por *M. massiliense* ser subestimado.

Neste contexto, a genômica comparativa se mostra uma ferramenta valiosa para a prospecção de alvos moleculares adequados para a discriminação das espécies do complexo

*M. abscessus*. Além disso, a análise do genoma fornece informações importantes sobre o funcionamento biológico e os mecanismos de virulência do organismo estudado, essenciais no desenvolvimento de novas estratégias para o tratamento eficaz de suas infecções e no controle de surtos.

## 5.1 Montagem e anotação

A partir da construção de um *pipeline* de análise computacional composto por diversas ferramentas de bioinformática, foi possível montar o genoma completo de *M. massiliense* GO 06. A montagem *de novo* de genomas não é uma tarefa fácil, sendo considerada um problema para o qual não se conhecem soluções computacionais eficientes [83]. Desta forma, a primeira tentativa de montagem do genoma de *M. massiliense* GO 06 foi por meio da abordagem *mapping* do MIRA, onde o cromossomo de uma espécie relacionada (*M. abscessus* ATCC 19977, neste caso) é utilizado para guiar o processo de montagem. O cromossomo gerado por esta abordagem apresentava diversas regiões sem cobertura, sendo enxertos do genoma de referência, e, por isso, foi descartado. Como a montagem de um genoma depende de uma série de fatores, como o tipo de sequenciamento ou a ferramenta de montagem empregada, e não se queria correr o risco da inserção de sequências de outros organismos no genoma montado, a montagem final do genoma de *M. massiliense* GO 06 foi obtida por meio da comparação da abordagem *de novo* de diversos montadores (CAP3, MIRA e AMOS). A montagem feita pelo MIRA se mostrou a mais adequada, e por isso foi escolhida. As sequências dos plasmídeos montados só puderam ser obtidas a partir da aplicação da metodologia desenvolvida em colaboração com o Departamento de Ciência da Computação da Universidade de Brasília.

O tamanho do cromossomo montado, de aproximadamente 4,7 Mb, está dentro do esperado para bactérias do gênero *Mycobacterium*, cujos genomas alcançam entre 3,3 Mb a 7 Mb, aproximadamente [84]. O contéudo de GC do genoma também se mostrou bastante similar ao de outras micobactérias (aproximadamente 65%), sendo coerente com o descrito para a espécie por HPLC [16]. A partir do GC *skew* do cromossomo de *M. massiliense* GO 06, é possível inferir os sítios de origem e término de replicação, representados pelas regiões onde existe a inversão do *skew*. Nos plasmídeos, o GC *skew* irregular sugere a presença de regiões adquiridas através da transferência horizontal de genes plasmidiais [85].

A grande similaridade (99% de identidade) do plasmídeo I com o plasmídeo pMAB01

de *M. abscessus* subp. *bolletii*, recentemente caracterizado *in vitro* [86], sugere a existência de pelo menos um plasmídeo no genoma de *M. massiliense* GO 06. Apesar de não apresentar similaridade com nenhuma sequência plasmidial micobacteriana disponível no repositório do NCBI, o plasmídeo II é uma sequência contígua e circular, cujo tamanho é consistente com o reportado para essas moléculas, podendo representar um plasmídeo ainda não caracterizado. Existe ainda a possibilidade de o plasmídeo II ser fruto de erros na montagem do genoma de *M. massiliense* GO 06. Mesmo havendo evidências da presença de plasmídeos no genoma desta estirpe, é necessário ainda caracterizá-los experimentalmente para confirmar a sua existência.

As análises *dot plot* são de grande utilidade para demonstrar a similaridade entre genomas de espécies relacionadas. A grande similaridade entre *M. massiliense* e *M. abscessus*, por exemplo, corrobora a consistência do genoma montado. Pode-se também observar a maior similaridade entre *M. massiliense* e *M. smegmatis*, uma espécie considerada saprofítica, em comparação a *M. tuberculosis* e *M. bovis*, duas bactérias patogênicas do complexo *M. tuberculosis*. Dentre as micobactérias analisadas, a que demonstrou menor similaridade a *M. massiliense* foi *M. leprae*. A baixa similaridade entre os genomas destas duas espécies pode ser explicado pelo alto grau de degeneração do genoma do agente etiológico da hanseníase [87]. Apesar de ser uma ferramenta interessante na visualização de regiões de similaridade entre os genomas, os *dot plots* não apresentam uma medida quantitativa da similaridade das sequências, ou seja, não existe uma significância estatística que possa ser testada.

O número de ORFs identificadas foi consistente com a quantidade presente em *M. abscessus*, sendo que a maior parte das ORFs preditas puderam ser anotadas. O número substancial de ORFs não anotadas pode ser explicado, em parte, pela limitação da abordagem de anotação por similaridade, onde a qualidade final da anotação é dependente da utilização de um banco de dados bem caracterizado. O desenvolvimento de um *pipeline* capaz de cruzar dados de anotações provenientes de alinhamentos contra bancos diferentes trouxe melhoras significativas na anotação do genoma de *M. massiliense* GO 06, em particular na anotação dos plasmídeos. O cuidado na anotação, tanto pela utilização de sequências provenientes de bancos curados, quanto no desenvolvimento de uma metodologia para comparação de fontes de anotação distintas é essencial para evitar erros de anotação e sua consequente propagação.

Por meio da categorização das ORFs nas classes do COG e da identificação dos fatores de virulência, foi possível traçar um perfil do funcionamento e da patogenicidade de *M.*

*massiliense* GO 06. A grande quantidade de genes relacionados à síntese da parede celular, por exemplo, sugere que a viabilidade celular de *M. massiliense* é tão dependente desta estrutura quanto a de *M. tuberculosis*.

Vale ressaltar, porém, que a definição de o que é um fator de virulência ainda é um tanto quanto controversa. Genes relacionados à manutenção da viabilidade de uma bactéria, podem, por exemplo, ser considerados fatores de virulência, mesmo não exercendo influência direta na patogenicidade do organismo. Desta forma, é necessário uma certa cautela

## 5.2 Sideróforos

Na descrição inicial de *M. massiliense* feita por Adékambi *et al.*, esta espécie não apresentou atividade de captação de ferro [16]. Entretanto, devido à importância desta atividade na patogenicidade de *M. tuberculosis*, as vias de biossíntese de sideróforos são de interesse no estudo da virulência de *M. massiliense*.

Existem evidências que *M. massiliense* sintetiza pelo menos um sideróforo com esta função biológica. Além da micobactina, já caracterizada como o principal sideróforo intracelular em micobactérias, observa-se que todas as enzimas relacionadas às vias de produção de enteroquelina, mixoquelina e bacilibactina puderam ser identificadas no genoma de *M. massiliense* GO 06. O gene *entA*, responsável pela síntese do precursor destas três moléculas (2,3-dihidroxibenzoato), é o mais expressivo entre os identificados para a via de sideróforos (aproximadamente 42,4% de todos os genes desta via).

Apesar da existência de evidências *in silico* da presença de genes que associados aos sideróforos no genoma de *M. massiliense* GO 06, a confirmação da expressão destes genes por ensaios laboratoriais ainda se faz necessária, particularmente por dois motivos: (1) um mesmo gene pode fazer parte de mais de uma via metabólica. O gene *entA*, por exemplo, além de participar na síntese de diversos sideróforos, também participa no metabolismo de metabólitos secundários relacionados a ácidos fenólicos, o que não significa, necessariamente, que as duas vias façam parte do metabolismo do organismo estudado; e (2) a anotação por similaridade pode ser ambígua quando um gene apresenta domínios conservados. O próprio *entA* apresenta domínios conservados relacionados a desidrogenases de cadeia curta. As vias metabólicas de sideróforos não ribossomais disponíveis no KEGG para *M. tuberculosis* indicam que este organismo sintetiza apenas a micobactina como sideróforo. Além dos genes da via de síntese de micobactina, as micobactérias do

complexo *M. abscessus* ainda apresentam o gene *entA*, não apresentando, porém, qualquer outro gene das vias de síntese dos outros sideróforos.

Além da micobactina e dos sideróforos externos, a carboximicobactina e a exoquelina, não existem relatos na literatura da presença de outras moléculas quelantes de ferro nas micobactérias. A carboximicobactina é uma variação da micobactina em que a porção da molécula responsável pela insolubilidade desta em meio aquoso é substituída. Embora as duas moléculas sejam bastante similares, a micobactina e a carboximicobactina não são interconversíveis, sendo provavelmente sintetizadas a partir de um precursor comum [88]. A carboximicobactina é característica de micobactérias patogênicas, apesar de presente em pequenas quantidades em organismos saprofíticos, não havendo nenhum relato da sua presença em micobactérias do complexo *M. abscessus*. A via de síntese da exoquelina, considerada o principal sideróforo externo de micobactérias saprofíticas, foi descrita em *M. smegmatis* [89]. Os principais genes desta via, *FxuA*, *FxuB* e *FxuC* são homólogos aos genes *FepG*, *FepC* e *FepD*, respectivamente, envolvidos na captação de ferro em *E. coli* [21]. O gene *FepC* foi identificado no genoma de *M. massiliense* GO 06, o que sugere a síntese de exoquelina por este organismo.

## 5.3 Sistemas de secreção bacterianos

Os sistemas de secreção são mecanismos de exportação de produtos bacterianos, necessários em diversas bactérias para a viabilidade e virulência do organismo. Enquanto os sistemas de secreção Sec-SRP e Tat são descritos tanto em bactérias Gram-positivas quanto nas Gram-negativas, os sistemas de secreção alternativos são característicos das últimas. Mesmo sendo classificadas como Gram-positivas, a parede celular das micobactérias apresenta um nível de complexidade similar ao de bactérias Gram-negativas. Desta forma, tanto as vias dos sistemas de secreção Sec-SRP e Tat quanto as vias dos sistemas alternativos são de interesse de estudo na descrição dos mecanismos de virulência de *M. massiliense*.

Apesar da identificação de genes envolvidos na biossíntese de certos componentes dos sistemas de secreção do tipo I, III, IV e VI, é possível observar que as vias de síntese desses mecanismos de secreção se encontram bastante incompletas no genoma de *M. massiliense* GO 06. Deste modo, é improvável que os sistemas de secreção alternativos sejam sintetizados por *M. massiliense*, o que é consistente com a literatura, já que não existem relatos da presença destes mecanismos em micobactérias.

A identificação de genes relacionados aos sistemas de secreção alternativos pode ser explicada por erros inerentes à anotação por similaridade de sequências. O gene *hlyB* do T1SS, por exemplo, o mais expressivo dentre os genes dos sistemas de secreção identificados no genoma de *M. massiliense* GO 06 (representando 49,5% de todos os genes desta via), está envolvido na síntese de um transportador ABC. Esses transportadores transmembrana são bastante comuns e estão envolvidos em uma série de funções biológicas, e sua presença no genoma não está necessariamente relacionada à existência do T1SS. Não é difícil imaginar que os vários transportadores ABC de uma célula bacteriana apresentam domínios bastante conservados, que podem levar a anotações ambíguas.

Ao contrário dos sistemas de secreção alternativos, os sistemas de secreção Sec-SRP e Tat estão descritos em micobactérias [25]. Com exceção de *secB*, *secG* e *secM*, todos os genes relacionados ao mecanismo de secreção Sec-SRP foram identificados no genoma de *M. massiliense* GO 06. A ausência de *secB* é esperada de acordo com a literatura, pois esta chaperona não é característica de bactérias Gram-positivas, como as micobactérias [25]. Curiosamente, o único gene identificado relacionado ao sistema de secreção Tat, *TatB*, é ausente na maioria das bactérias Gram-positivas [90]. Nestas bactérias, *TatA* é uma proteína bifuncional, apresentando a função de *TatA* e *TatB* das Gram-negativas. Desta forma, é possível que o gene identificado em *M. massiliense* seja na verdade o *TatA* bifuncional, que apresenta regiões homólogas ao *TatB*.

## 5.4 Sistema de secreção do tipo VII

Sistemas homólogos aos sistemas de secreção alternativos são ausentes no genoma de *M. tuberculosis*, entretanto, um sistema funcionalmente equivalente, denominado ESX-1 (e, mais tarde, sistema de secreção do tipo VII), foi caracterizado nesta espécie. O genoma de *M. tuberculosis* contém ainda outros quatro *loci* com genes homólogos aos do ESX-1, nomeados ESX-2 a ESX-5 [26]. Esses sistemas são ainda pouco estudados, mas sabe-se que apesar de ESX-1 ser essencial para a virulência de *M. tuberculosis*, nem todos os T7SS estão relacionados à patogenicidade do organismo [26]. Sabe-se também que os substratos secretados por cada sistema podem diferir [25, 91].

Em *M. massiliense* GO 06, foram identificados dois *loci* de genes associados aos T7SS. No primeiro *locus*, todos os genes ESX apresentam maior similaridade ao ESX-3 de *M. tuberculosis*. Já no segundo *locus*, os genes ESX identificados apresentam maior similaridade aos genes do ESX-4 de *M. tuberculosis*, com exceção do EccE, que apresenta

maior similaridade ao EccE5 (ESX-5). Também foram identificados genes das famílias ESAT-6 e FtsK/SpoIIIE, associados ao T7SS, em outras regiões genômicas.

A ausência de genes relacionados ao ESX-1 pode explicar o fato de *M. massiliense* ser uma bactéria oportunista, ao contrário de *M. tuberculosis*. O sistema ESX-4 é considerado o T7SS mais ancestral em micobactérias, e os seus substratos de secreção ainda não foram detectados em meio extracelular [26]. Em contraste, as proteínas da famíla ESAT-6 secretadas por ESX-3 foram identificadas no meio extracelular, caracterizando este sistema como um mecanismo de secreção funcional. Apesar do papel de ESX-4 na virulência das micobactérias não ser claro, foi demonstrado que o ESX-3 tem papel essencial na via da aquisição de ferro por meio da micobactina [92] e é necessário para o crescimento em cultura de *M. tuberculosis* [26].

## 5.5 Divulgação dos dados de anotação

Embora crescimento de surtos relacionados às micobactérias de crescimento rápido, a quantidade de informações biológicas depositadas em bancos de dados referentes a esses organismos ainda é muito pequena. A disponibilização dos dados de anotação e montagem do genoma de *M. massiliense* GO 06 no repositório do NCBI, representando o primeiro genoma completo de *M. massiliense*, e por meio de uma página própria hospedada no domínio da Universidade de Brasília é uma grande contribuição científica em diversas áreas do conhecimento. As análises *in silico* dos fatores de virulência e vias metabólicas de *M. massiliense* GO 06 também são de grande importância científica, pois, além de contribuirem para o entendimento sobre o funcionamento deste organismo, também fornecem informações para novos estudos de identificação de alvos moleculares para a identificação de estirpes ou para a atenuação da patogenicidade do organismo.

# 6 Conclusão

Foi possível, por meio de um *pipeline* computacinal, montar o genoma completo de *M. massiliense* GO 06, formado por seu cromossomo e dos plasmídeos. A maior parte das ORFs identificadas no genoma desta estirpe pôde ser anotada, bem como 46 tRNAs e um operon de rRNA.

No total, 826 genes puderam ser relacionados a fatores de virulência, a maioria destes relacionados ao metabolismo de lipídios e à biossíntese de sideróforos e sistemas de secreção bacterianos. A partir da identificação dos genes de virulência de *M. massiliense*, foi possível traçar o perfil de virulência deste organismo, bem como verificar a existência de mecanismos de patogenicidade importantes. As vias metabólicas relacionadas aos sistemas de secreção bacterianos e à biossíntese de sideróforos, importantes na virulência de micobactérias, foram caracterizadas *in silico*. Além disso, dois operons de genes envolvidos na biossíntese do T7SS, homólogos ao ESX-3 e ESX-4 de *M. tuberculosis*, foram identificados no genoma de *M. massiliense* GO 06.

Os dados *in silico* referentes ao genoma do isolado GO 06 representam um conhecimento científico valioso que pode ser útil no desenvolvimento métodos de identificação das espécies do complexo *M. abscessus* e de estratégias para o controle de surtos relacionados a *M. massiliense*.

# *Referências*


[1] National Institute of Allergy and Infectious Diseases. `http://www.niaid.nih.gov/LabsAndResources/resources/translational/microscopy/Pages/sem.aspx`. Acessado em 17 de Janeiro de 2014.

[2] Pathogen Profile Dictionary: *Mycobacterium leprae*. `http://www.ppdictionary.com/bacteria/gpbac/leprae.htm`. Acessado em 17 de Janeiro de 2014.

[3] MADIGAN, M.; MARTINKO, J. Brock biology of microorganisms. Pearson Prentice Hall, 2006.

[4] MADISON, B. Application of stains in clinical microbiology. Biotechnic & Histochemistry, v. 76, n. 3, p. 119–125, 2001.

[5] RYAN, K.; RAY, C. Sherris medical microbiology. Lange Basic Science. Mcgraw-hill, 2003.

[6] University of Toronto: Department of Molecular Genetics. `http://www.utoronto.ca/liulab/Projects.html`. Acessado em 17 de Janeiro de 2014.

[7] DENYER, S.; MAILLARD, J.-Y. Cellular impermeability and uptake of biocides and antibiotics in gram-negative bacteria. Journal of applied microbiology, v. 92, n. s1, p. 35S–45S, 2002.

[8] HAN, X. Y.; DÉ, I.; JACOBSON, K. L. Rapidly growing mycobacteria clinical and microbiologic studies of 115 cases. American journal of clinical pathology, v. 128, n. 4, p. 612–621, 2007.

[9] DE GROOTE, M. A.; HUITT, G. Infections due to rapidly growing mycobacteria. Clinical infectious diseases, v. 42, n. 12, p. 1756–1763, 2006.

[10] HALL-STOODLEY, L.; STOODLEY, P. Biofilm formation and dispersal and the transmission of human pathogens. Trends in microbiology, v. 13, n. 1, p. 7–10, 2005.

[11] OMS. Doenças causadas por micobactérias não tuberculosas. `http://apps.who.int/medicinedocs/en/d/Js5511e/4.html`. Acessado em 21 de Janeiro de 2014.

[12] BARON, S. Medical microbiology. Churchill Livingstone, 1991.

[13] CARSON, L. A.; PETERSEN, N. J.; FAVERO, M. S.; AGUERO, S. Growth characteristics of atypical mycobacteria in water and their comparative resistance to disinfectants. Applied and environmental microbiology, v. 36, n. 6, p. 839–846, 1978.

[14] CARDOSO, A. M.; JUNQUEIRA-KIPNIS, A. P.; KIPNIS, A. *In vitro* antimicrobial susceptibility of *Mycobacterium massiliense* recovered from wound samples of patients submitted to arthroscopic and laparoscopic surgeries. Minimally invasive surgery, v. 2011, 2011.

[15] EID, A. J.; BERBARI, E. F.; SIA, I. G.; WENGENACK, N. L.; OSMON, D. R.; RAZONABLE, R. R. Prosthetic joint infection due to rapidly growing mycobacteria: report of 8 cases and review of the literature. Clinical Infectious Diseases, v. 45, n. 6, p. 687–694, 2007.

[16] ADÉKAMBI, T.; REYNAUD-GAUBERT, M.; GREUB, G.; GEVAUDAN, M.-J.; LA SCOLA, B.; RAOULT, D.; DRANCOURT, M. Amoebal coculture of "*Mycobacterium massiliense*" sp. nov. from the sputum of a patient with hemoptoic pneumonia. Journal of clinical microbiology, v. 42, n. 12, p. 5493–5501, 2004.

[17] ZELAZNY, A. M.; ROOT, J. M.; SHEA, Y. R.; COLOMBO, R. E.; SHAMPUTA, I. C.; STOCK, F.; CONLAN, S.; MCNULTY, S.; BROWN-ELLIOTT, B. A.; WALLACE, R. J. et al. Cohort study of molecular identification and typing of *Mycobacterium abscessus*, *Mycobacterium massiliense*, and *Mycobacterium bolletii*. Journal of clinical microbiology, v. 47, n. 7, p. 1985–1995, 2009.

[18] LEAO, S. C.; TORTOLI, E.; VIANA-NIERO, C.; UEKI, S. Y. M.; LIMA, K. V. B.; LOPES, M. L.; YUBERO, J.; MENENDEZ, M. C.; GARCIA, M. J. Characterization of mycobacteria from a major brazilian outbreak suggests that revision of the taxonomic status of members of the *Mycobacterium chelonae-M. abscessus* group is needed. Journal of clinical microbiology, v. 47, n. 9, p. 2691–2698, 2009.

[19] SIMMON, K. E.; POUNDER, J. I.; GREENE, J. N.; WALSH, F.; ANDERSON, C. M.; COHEN, S.; PETTI, C. A. Identification of an emerging pathogen, *Mycobacterium massiliense*, by rpoB sequencing of clinical isolates collected in the united states. Journal of clinical microbiology, v. 45, n. 6, p. 1978–1980, 2007.

[20] KIM, H.-Y.; KOOK, Y.; YUN, Y.-J.; PARK, C. G.; LEE, N. Y.; SHIM, T. S.; KIM, B.-J.; KOOK, Y.-H. Proportions of *Mycobacterium massiliense* and *Mycobacterium bolletii* strains among korean *Mycobacterium chelonae-Mycobacterium abscessus* group isolates. Journal of clinical microbiology, v. 46, n. 10, p. 3384–3390, 2008.

[21] RATLEDGE, C. Iron, mycobacteria and tuberculosis. Tuberculosis, v. 84, n. 1, p. 110–130, 2004.

[22] RODRIGUEZ, G. M. Control of iron metabolism in *Mycobacterium tuberculosis*. TRENDS in Microbiology, v. 14, n. 7, p. 320–327, 2006.

[23] DAFFÉ, M.; DRAPER, P. The envelope layers of mycobacteria with reference to their pathogenicity. Advances in microbial physiology, v. 39, p. 131–203, 1997.

[24] FUKUDA, T.; MATSUMURA, T.; ATO, M.; HAMASAKI, M.; NISHIUCHI, Y.; MURAKAMI, Y.; MAEDA, Y.; YOSHIMORI, T.; MATSUMOTO, S.; KOBAYASHI, K. et al. Critical roles for lipomannan and lipoarabinomannan in cell wall integrity of mycobacteria and pathogenesis of tuberculosis. MBio, v. 4, n. 1, p. e00472–12, 2013.

[25] DIGIUSEPPE CHAMPION, P. A.; COX, J. S. Protein secretion systems in mycobacteria. Cellular microbiology, v. 9, n. 6, p. 1376–1384, 2007.

[26] ABDALLAH, A. M.; VAN PITTIUS, N. C. G.; CHAMPION, P. A. D.; COX, J.; LUIRINK, J.; VANDENBROUCKE-GRAULS, C. M.; APPELMELK, B. J.; BITTER, W. Type VII secretion - mycobacteria show the way. Nature reviews microbiology, v. 5, n. 11, p. 883–891, 2007.

[27] VAN PITTIUS, N. G.; GAMIELDIEN, J.; HIDE, W.; BROWN, G. D.; SIEZEN, R. J.; BEYERS, A. D. The ESAT-6 gene cluster of *Mycobacterium tuberculosis* and other high G+C Gram-positive bacteria. Genome Biol, v. 2, n. 10, p. 44–1, 2001.

[28] STANLEY, S. A.; RAGHAVAN, S.; HWANG, W. W.; COX, J. S. Acute infection and macrophage subversion by *Mycobacterium tuberculosis* require a specialized secretion system. Proceedings of the National Academy of Sciences, v. 100, n. 22, p. 13001–13006, 2003.

[29] GUINN, K. M.; HICKEY, M. J.; MATHUR, S. K.; ZAKEL, K. L.; GROTZKE, J. E.; LEWINSOHN, D. M.; SMITH, S.; SHERMAN, D. R. Individual RD1-region genes are required for export of ESAT-6/CFP-10 and for virulence of *Mycobacterium tuberculosis*. Molecular microbiology, v. 51, n. 2, p. 359–370, 2004.

[30] PETRINI, B. *Mycobacterium abscessus*: an emerging rapid-growing potential pathogen. Apmis, v. 114, n. 5, p. 319–328, 2006.

[31] KOH, W.-J.; JEON, K.; LEE, N. Y.; KIM, B.-J.; KOOK, Y.-H.; LEE, S.-H.; PARK, Y. K.; KIM, C. K.; SHIN, S. J.; HUITT, G. A. et al. Clinical significance of differentiation of *Mycobacterium massiliense* from *Mycobacterium abscessus*. American Journal of Respiratory and Critical Care Medicine, v. 183, n. 3, p. 405–410, 2011.

[32] VIANA-NIERO, C.; LIMA, K. V. B.; LOPES, M. L.; DA SILVA RABELLO, M. C.; MARSOLA, L. R.; BRILHANTE, V. C. R.; DURHAM, A. M.; LEÃO, S. C. Molecular characterization of *Mycobacterium massiliense* and *Mycobacterium bolletii* in isolates collected from outbreaks of infections after laparoscopic surgeries and cosmetic procedures. Journal of clinical microbiology, v. 46, n. 3, p. 850–855, 2008.

[33] CARDOSO, A. M.; MARTINS DE SOUSA, E.; VIANA-NIERO, C.; BONFIM DE BORTOLI, F.; PEREIRA DAS NEVES, Z. C.; LEÃO, S. C.; JUNQUEIRA-KIPNIS, A. P.; KIPNIS, A. Emergence of nosocomial *Mycobacterium massiliense* infection in goiás, brazil. Microbes and Infection, v. 10, n. 14, p. 1552–1557, 2008.

[34] DUARTE, R. S.; LOURENÇO, M. C. S.; DE SOUZA FONSECA, L.; LEAO, S. C.; EFIGENIA DE LOURDES, T. A.; ROCHA, I. L.; COELHO, F. S.; VIANA-NIERO, C.; GOMES, K. M.; DA SILVA, M. G. et al. Epidemic of postsurgical infections caused by *Mycobacterium massiliense*. Journal of clinical microbiology, v. 47, n. 7, p. 2149–2155, 2009.

[35] FLEISCHMANN, R. D.; ADAMS, M. D.; WHITE, O.; CLAYTON, R. A.; KIRKNESS, E. F.; KERLAVAGE, A. R.; BULT, C. J.; TOMB, J.-F.; DOUGHERTY, B. A.; MERRICK, J. M. et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* rd. Science, Washington, v. 269, n. 5223, p. 496–512, 1995.

[36] FRASER, C. M.; GOCAYNE, J. D.; WHITE, O.; ADAMS, M. D.; CLAYTON, R. A.; FLEISCHMANN, R. D.; BULT, C. J.; KERLAVAGE, A. R.; SUTTON, G.; KELLEY, J. M. et al. The minimal gene complement of *Mycoplasma genitalium*. Science, Washington, v. 270, n. 5235, p. 397–404, 1995.

[37] KOONIN, E. V.; WOLF, Y. I. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic acids research, v. 36, n. 21, p. 6688–6719, 2008.

[38] OSTERMAN, A.; OVERBEEK, R. Missing genes in metabolic pathways: a comparative genomics approach. Current opinion in chemical biology, v. 7, n. 2, p. 238–251, 2003.

[39] FONG, S. S.; BURGARD, A. P.; HERRING, C. D.; KNIGHT, E. M.; BLATTNER, F. R.; MARANAS, C. D.; PALSSON, B. O. *In silico* design and adaptive evolution of *Escherichia coli* for production of lactic acid. Biotechnology and bioengineering, v. 91, n. 5, p. 643–648, 2005.

[40] PHARKYA, P.; BURGARD, A. P.; MARANAS, C. D. OptStrain: a computational framework for redesign of microbial production systems. Genome research, v. 14, n. 11, p. 2367–2376, 2004.

[41] LOCKHART, D. J.; WINZELER, E. A. Genomics, gene expression and DNA arrays. nature, v. 405, n. 6788, p. 827–836, 2000.

[42] BUYSSE, J. M. The role of genomics in antibacterial target discovery. Current medicinal chemistry, v. 8, n. 14, p. 1713–1726, 2001.

[43] BUTLER, G.; RASMUSSEN, M. D.; LIN, M. F.; SANTOS, M. A.; SAKTHIKUMAR, S.; MUNRO, C. A.; RHEINBAY, E.; GRABHERR, M.; FORCHE, A.; REEDY, J. L. et al. Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes. Nature, London, v. 459, n. 7247, p. 657–662, 2009.

[44] COLE, S.; BROSCH, R.; PARKHILL, J.; GARNIER, T.; CHURCHER, C.; HARRIS, D.; GORDON, S.; EIGLMEIER, K.; GAS, S.; BARRY, C. R. et al. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature, London, v. 393, n. 6685, p. 537–544, 1998.

[45] ZHANG, Y.-Q.; REN, S.-X.; LI, H.-L.; WANG, Y.-X.; FU, G.; YANG, J.; QIN, Z.-Q.; MIAO, Y.-G.; WANG, W.-Y.; CHEN, R.-S. et al. Genome-based analysis of virulence genes in a non-biofilm-forming *Staphylococcus epidermidis* strain (atcc 12228). Molecular microbiology, v. 49, n. 6, p. 1577–1593, 2003.

[46] SAKHARKAR, K. R.; SAKHARKAR, M. K.; CHOW, V. T. A novel genomics approach for the identification of drug targets in pathogens, with special reference to *Pseudomonas aeruginosa*. In silico biology, v. 4, n. 3, p. 355–360, 2004.

[47] Genomes Online Database. `http://genomesonline.org/cgi-bin/GOLD/index.cgi`. Acessado em 19 de Janeiro de 2014.

[48] NCBI. NCBI-Genbank flat file release 199. `ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt`, 2013. Acessado em 19 de Janeiro de 2014.

[49] SANGER, F.; NICKLEN, S.; COULSON, A. R. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences, v. 74, n. 12, p. 5463–5467, 1977.

[50] MOROZOVA, O.; MARRA, M. A. Applications of next-generation sequencing technologies in functional genomics. Genomics, v. 92, n. 5, p. 255–264, 2008.

[51] HALL, N. Advanced sequencing technologies and their wider impact in microbiology. Journal of Experimental Biology, v. 210, n. 9, p. 1518–1525, 2007.

[52] LIU, L.; LI, Y.; LI, S.; HU, N.; HE, Y.; PONG, R.; LIN, D.; LU, L.; LAW, M. Comparison of next-generation sequencing systems. Journal of Biomedicine and Biotechnology, v. 2012, 2012.

[53] MARGULIES, M.; EGHOLM, M.; ALTMAN, W. E.; ATTIYA, S.; BADER, J. S.; BEMBEN, L. A.; BERKA, J.; BRAVERMAN, M. S.; CHEN, Y.-J.; CHEN, Z. et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature, London, v. 437, n. 7057, p. 376–380, 2005.

[54] METZKER, M. L. Sequencing technologies - the next generation. Nature Reviews Genetics, v. 11, n. 1, p. 31–46, 2009.

[55] CHOU, H.-H.; HOLMES, M. H. Dna sequence quality trimming and vector removal. Bioinformatics, v. 17, n. 12, p. 1093–1104, 2001.

[56] Phrap: Phred Quality Base Calling. `http://www.phrap.com/phred/`. Acessado em 19 de Janeiro de 2014.

[57] Página do AMOS. `http://sourceforge.net/apps/mediawiki/amos/index.php?title=AMOS`. Acessado em 27 de Janeiro de 2014.

[58] CHEVREUX, B. MIRA: an automated genome and est assembler. Ruprecht-Karls University, Heidelberg, Germany, 2005.

[59] PHAST. `http://gcat.davidson.edu/phast/`. Acessado em 19 de Janeiro de 2014.

[60] YANDELL, M.; ENCE, D. A beginner's guide to eukaryotic genome annotation. Nature Reviews Genetics, v. 13, n. 5, p. 329–342, 2012.

[61] PEVSNER, J. Bioinformatics and functional genomics. Wiley, 2009.

[62] DELCHER, A. L.; HARMON, D.; KASIF, S.; WHITE, O.; SALZBERG, S. L. Improved microbial gene identification with GLIMMER. Nucleic acids research, v. 27, n. 23, p. 4636–4641, 1999.

[63] WARREN, A. S.; SETUBAL, J. C. The Genome Reverse Compiler: an explorative annotation tool. BMC bioinformatics, v. 10, n. 1, p. 35, 2009.

[64] University of Pittsburgh: Alignment. `http://www.pitt.edu/~mcs2/teaching/biocomp/tutorials/global.html`. Acessado em 19 de Janeiro de 2014.

[65] ALTSCHUL, S. F.; MADDEN, T. L.; SCHÄFFER, A. A.; ZHANG, J.; ZHANG, Z.; MILLER, W.; LIPMAN, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research, v. 25, n. 17, p. 3389–3402, 1997.

[66] Babraham Institute: FastQC. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Acessado em 21 de Janeiro de 2014.

[67] PHAST: *Overlap Layout Consensus*. http://gcat.davidson.edu/phast/olc.html. Acessado em 19 de Janeiro de 2014.

[68] University of Leipzig: Segemehl. http://www.bioinf.uni-leipzig.de/Software/segemehl/. Acessado em 25 de Janeiro de 2014.

[69] GenBank. http://www.ncbi.nlm.nih.gov/genbank/. Acessado em 21 de Janeiro de 2014.

[70] Cluster of Orthologous Groups. http://www.ncbi.nlm.nih.gov/COG/. Acessado em 23 de Janeiro de 2014.

[71] Kyoto Encyclopedia of Genes and Genomes. http://www.genome.jp/kegg/. Acessado em 22 de Janeiro de 2014.

[72] UniProt Knowledgebase. http://www.uniprot.org/help/uniprotkb. Acessado em 23 de Janeiro de 2014.

[73] Virulence Factor Database. http://www.mgc.ac.cn/VFs/. Acessado em 21 de Janeiro de 2014.

[74] COMAV Institute: sff_extract. http://bioinf.comav.upv.es/sff_extract/. Acessado em 23 de Janeiro de 2014.

[75] Página do MUMmer. http://mummer.sourceforge.net/. Acessado em 28 de Janeiro de 2014.

[76] LAGESEN, K.; HALLIN, P.; RØDLAND, E. A.; STÆRFELDT, H.-H.; ROGNES, T.; USSERY, D. W. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic acids research, v. 35, n. 9, p. 3100–3108, 2007.

[77] SCHATTNER, P.; BROOKS, A. N.; LOWE, T. M. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic acids research, v. 33, n. suppl 2, p. W686–W689, 2005.

[78] RIBEIRO, G. M.; RAIOL, T.; MARANHÃO, A. Q.; SILVA-PEREIRA, I.; BOCCA, A. L.; JUNQUEIRA-KIPNIS, A. P.; WALTER, M. E. M. et al. Genome sequence of Mycobacterium massiliense strain go 06 and comparative genomics analysis. In: BSB 2012 Digital Proceedings. 2012.

[79] MENEGÓI, G.; RAIOL, T.; DE ARAÚJO OLIVEIRA, J. V.; DE OLIVEIRA SANDES, E. F.; DE MELO, A. C. M. A.; MARANHÃO, A. Q.; SILVA-PEREIRA, I.; BOCCA, A. L.; JUNQUEIRA-KIPNIS, A. P.; WALTER, M. E. M. et al. A pipeline to characterize virulence factors in Mycobacterium Massiliense genome. In: Advances in Bioinformatics and Computational Biology. Springer, 2013. p. 202–213.

[80] RAIOL, T.; RIBEIRO, G. M.; MARANHÃO, A. Q.; BOCCA, A. L.; SILVA-PEREIRA, I.; JUNQUEIRA-KIPNIS, A. P.; DE MACEDO BRIGIDO, M.; KIPNIS, A. Complete genome sequence of Mycobacterium massiliense. Journal of bacteriology, v. 194, n. 19, p. 5455–5455, 2012.

[81] MACHERAS, E.; ROUX, A.-L.; RIPOLL, F.; SIVADON-TARDY, V.; GUTIERREZ, C.; GAILLARD, J.-L.; HEYM, B. Inaccuracy of single-target sequencing for discriminating species of the *Mycobacterium abscessus* group. Journal of clinical microbiology, v. 47, n. 8, p. 2596–2600, 2009.

[82] JANDA, J. M.; ABBOTT, S. L. 16s rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. Journal of clinical microbiology, v. 45, n. 9, p. 2761–2764, 2007.

[83] POP, M. Genome assembly reborn: recent computational challenges. Briefings in bioinformatics, v. 10, n. 4, p. 354–366, 2009.

[84] TB Database: genome statistics. `http://genome.tbdb.org/annotation/genome/tbdb/GenomeStats.html`. Acessado em 12 de Fevereiro de 2014.

[85] SOARES, S. C.; ABREU, V. A.; RAMOS, R. T.; CERDEIRA, L.; SILVA, A.; BAUMBACH, J.; TROST, E.; TAUCH, A.; HIRATA JR, R.; MATTOS-GUARALDI, A. L. et al. PIPS: pathogenicity island prediction software. PloS one, v. 7, n. 2, p. e30848, 2012.

[86] LEAO, S. C.; MATSUMOTO, C. K.; CARNEIRO, A.; RAMOS, R. T.; NOGUEIRA, C. L.; JUNIOR, J. D. L.; LIMA, K. V.; LOPES, M. L.; SCHNEIDER, H.; AZEVEDO, V. A. et al. The detection and sequencing of a broad-host-range conjugative incP-1$\beta$ plasmid in an epidemic strain of *Mycobacterium abscessus* subsp. *bolletii*. PloS one, v. 8, n. 4, p. e60746, 2013.

[87] COLE, S.; EIGLMEIER, K.; PARKHILL, J.; JAMES, K.; THOMSON, N.; WHEELER, P.; HONORE, N.; GARNIER, T.; CHURCHER, C.; HARRIS, D. et al. Massive gene decay in the leprosy bacillus. Nature, London, v. 409, n. 6823, p. 1007–1011, 2001.

[88] RATLEDGE, C.; DALE, J. Mycobacteria: Molecular biology and virulence. Wiley, 2009.

[89] ZHU, W.; ARCENEAUX, J. E.; BEGGS, M. L.; BYERS, B. R.; EISENACH, K. D.; LUNDRIGAN, M. D. Exochelin genes in *Mycobacterium smegmatis*: identification of an abc transporter and two non-ribosomal peptide synthetase genes. Molecular microbiology, v. 29, n. 2, p. 629–639, 1998.

[90] BARNETT, J. P.; EIJLANDER, R. T.; KUIPERS, O. P.; ROBINSON, C. A minimal Tat system from a Gram-positive organism a bifunctional TatA subunit participates in discrete TatAC and TatA complexes. Journal of biological chemistry, v. 283, n. 5, p. 2534–2542, 2008.

[91] ABDALLAH, A. M.; BESTEBROER, J.; SAVAGE, N. D.; DE PUNDER, K.; VAN ZON, M.; WILSON, L.; KORBEE, C. J.; VAN DER SAR, A. M.; OTTENHOFF,

T. H.; VAN DER WEL, N. N. et al. Mycobacterial secretion systems ESX-1 and ESX-5 play distinct roles in host cell death and inflammasome activation. The Journal of Immunology, v. 187, n. 9, p. 4744–4753, 2011.

[92] SIEGRIST, M. S.; UNNIKRISHNAN, M.; MCCONNELL, M. J.; BOROWSKY, M.; CHENG, T.-Y.; SIDDIQI, N.; FORTUNE, S. M.; MOODY, D. B.; RUBIN, E. J. Mycobacterial Esx-3 is required for mycobactin-mediated iron acquisition. Proceedings of the National Academy of Sciences, v. 106, n. 44, p. 18792–18797, 2009.

# *APÊNDICE A – Artigo apresentado no BSB 2012*

## Genome sequence of *Mycobacterium massiliense* strain GO 06 and comparative genomics analysis

Guilherme Menegói Ribeiro[1], Tainá Raiol[1], Andréa Queiroz Maranhão[1], Ildinete Silva Pereira[1], Anamélia Lorenzetti Bocca[1], Ana Paula Junqueira-Kipnis[2], Marcelo de Macedo Brígido[1], and André Kipnis[2]

[1] Department of Cellular Biology, Institute of Biology, University of Brasilia, 70910-900, Brasília, DF, Brazil
[2] Department of Microbiology, Immunology, Parasitology, and Pathology, Federal University of Goiás, 74605-050, Goiania, GO, Brazil

**Abstract.** The *Mycobacterium massiliense* strain GO 06 was isolated from wound samples of patients submitted to arthroscopic and laparoscopic surgeries. This strain had its entire genome sequenced using the 454 GS-FLX Titanium (Roche) high-throughput sequencer. It was possible to construct strain GO 06 entire chromosome and annotate the majority of conserved ORFs. In addition, similarity analysis showed that *M. massiliense* is much more related to pathogenic mycobacteria, particularly *Mycobacterium abscessus*.

**Keywords:** genome; *Mycobacterium massiliense*; bioinformatics; nosocomial infections

### 1 Introduction

The number of reported cases related to nontuberculous mycobacteria (NTM) has greatly increased during the last few years, following the improvement of culture and identification techniques. Among those bacteria, mycobacteria of the *Mycobacterium chelonae-Mycobacterium abscessus* group (*M. chelonae*, *M. abscessus*, and *Mycobacterium immunogenum*), particularly *M. abscessus*, arose as the most important opportunistic pathogens. However, in 2004, Adékambi *et al.* [1] assigned a novel species for a closely related isolate, *Mycobacterium massiliense*, a representative species of rapidly growing mycobacteria (RGM). RGM have important implications in human diseases, as they are frequently associated with infections among immunocompromised patients as well as wound, skin, and soft tissue infections [2]. Additionally, these bacteria are naturally resistant to several classes of antibiotics, and particularly to antituberculousis drugs.

Ever since its description, *M. massiliense* has been increasingly reported as causing soft tissue infection outbreaks. In Brazil, a major outbreak has been recently reported with the characterization of some aspects of antibiotic resistance as well as disinfectant resistance that may have also contributed to the difficulty in controlling the spread of this strain.

*M. massiliense* is a strictly aerobic, non-spore-forming, nonmotile, acid-fast, gram-positive rod that shares 100% of its 16S rRNA sequence with *Mycobacterium abscessus* [1].

Sequencing and annotation of *M. massiliense*'s genome have a crucial role in defining some of its virulence aspects, as well as its metabolic pathways and phylogenetic classification, which will in turn greatly impact our understanding on the mechanics of infection by *M. massiliense* and the evolution of its pathogenicity.

## 2 Methods

### 2.1 Genome sequencing and contig assembly

A nontuberculous mycobacteria, identified as *Mycobacterium massiliense* strain GO 06, by phenotypic and molecular methods, was isolated from a Brazilian patient who had undergone knee joint surgery [3]. Its genome was sequenced by a whole-genome shotgun approach using 454 GS-FLX Titanium (Roche) high-throughput sequencer. Contigs were constructed by MIRA assembler's mapping method, which uses a closely related species genome as a reference backbone for assembly. The reference for our assembly was the complete chromosome of *M. abscessus* str. ATCC 19977, available at NCBI (access number CU458896).

### 2.2 Comparative genomics analysis

To evaluate the similarity relatedness with other mycobacteria, a series of dotplots were constructed by mummerplot [4] comparing the genomes of some mycobacteria (*M. abscessus* str. ATCC 19977, *Mycobacterium leprae* str. Br4923 (NC_011896), *M. tuberculosis* str. CDC1551 (NC_002755) and *Mycobacterium smegmatis* str. MC2 155 (NC_008596)) to the assembled genome of strain GO 06. A graphical visualization of the chromosome and its GC content was made by CGView server (`http://stothard.afns.ualberta.ca/cgview_server/`).

### 2.3 ORF prediction and annotation

ORFs (open reading frames) were predicted by GRC (Genome Reverse Compiler), a prokaryotic annotation tool that finds all possible ORFs within a genome and then evaluates the likelihood of translation to a protein [5]. GRC selects ORFs that are likely to be a real gene and runs an annotation procedure utilizing a database reference of well annotated, closely related genomes. For this study, all GRC settings were used as default, and the list of genomes used for the annotation and their respective acession numbers are as follows: *M. abscessus* str. ATCC 19977, *M. bovis* str. AF2122/97 (NC_002945), *Mycobacterium bovis* str. Mexico (NC_016804), *M. bovis* str. Pasteur 1173P2 (NC_008769), *M. bovis* str. Tokyo 172 (NC_012207), *M. leprae* str. Br4923, *M. leprae* str. TN (NC_002677), *M. tuberculosis* str. H37Rv (NC_000962), *M. tuberculosis* str. CDC1551, *M. tuberculosis* str. H37Ra (NC_009525) and *M. tuberculosis* str. CDC5180 (NC_017522).

Additional information was also assigned to predicted ORFs by a BLASTp using a database constructed from all bacterial proteins from Swiss-Prot (`http://www.uniprot.org/uniprot/?query=reviewed\%3Ayes`) .

COG category was determined by BLASTp comparison between GO 06 ORFs and Clusters of Orthologous Groups of proteins (COGs) database (`http://www.ncbi.nlm.nih.gov/COG/`).

## 3 Results and Discussion

### 3.1 Genome assembly and annotation

Out of 584,619 reads, 446,501 (76.3%) were assembled into a single contig of 5,068,807 bp, corresponding to the *M. massiliense* chromosome. Its general features are listed in Table 1. As seen in Figure 1, GO 06 genome sequence presents the highest similarity with *M. abscessus*, being less similar to other mycobacteria, particularly *M. smegmatis*, which is generally considered to be non-pathogenic.

**Table 1.** General features of *M. massiliense* str. GO 06 chromosome.

| | |
|---|---|
| Chromosome (bp) | 5,068,807 |
| Chromosomal GC content (%) | 64.2% |
| Total number of contigs | 1 |
| Average contig size (bp) | 5,068,807 |
| Total number of ORFs | 4,313 |
| Number of conserved ORFs with known protein function | 2,311 |
| Number of conserved ORFs without assigned function | 906 |
| Number of nonconserved ORFs | 1,096 |

The GC content for the chromosome was similar to those reported for other mycobacteria (approximately 65%), also agreeing with the GC content determined by HPLC [1]. Figure 2 is a graphical representation of isolate GO 06's chromosome and its GC content, in comparison with other three mycobacteria (*M. abscessus* str. ATCC 19977, *M. bovis* str. Pasteur 1173P2 and *M. tuberculosis* str. CDC 1551).

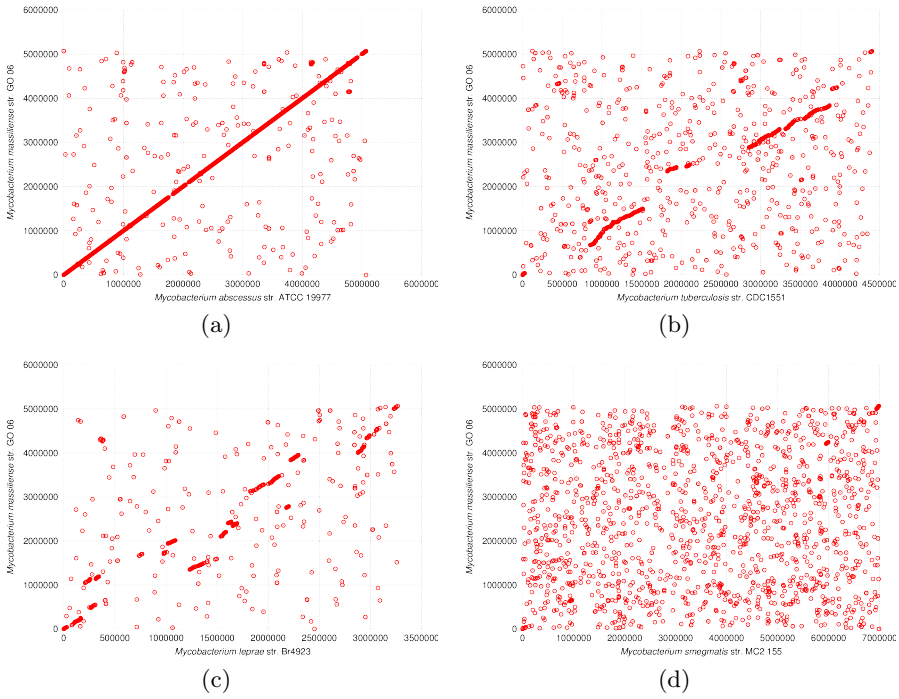



**Fig. 1.** Dotplot analysis between isolate GO 06 and other mycobacteria: (a) *M. abscessus* str. ATCC 19977, (b) *M. tuberculosis* str. CDC 1551, (c) *M. leprae* str. Br4923 and (d) *M. smegmatis* str. MC2 155.

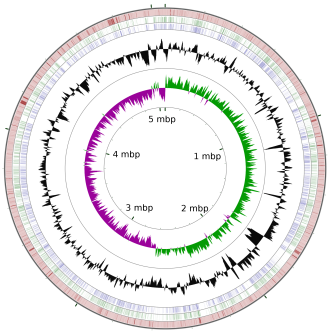



**Fig. 2.** Chromosome of isolate GO 06 and similarity comparison with other mycobacteria. Each circle, from the inner to outer circles, represents: GO 06 isolate's circular chromosome, GC skew (skew plus in green and skew minus in pink), GC content ratios and the similarity analysis in comparison to *M. abscessus* str. ATCC 19977 (red), *M. bovis* str. Pasteur 1173P2 (green) and *M. tuberculosis* str. CDC 1551 (blue). In the latter three circles, the color strength is directly proportional to the similarity score generated by Blast.

Out of 4,313 ORFs, 2,869 were assigned to COG categories, corresponding to 90% of conserved ORFs. Figure 3 shows the distribution in percentage of each COG category. Roughly 50% of all assigned categories were related to cell metabolism, 16% to information storage and processing and 11% to cellular processes and signaling, while 22% of all categories were in the "poorly characterized" class.

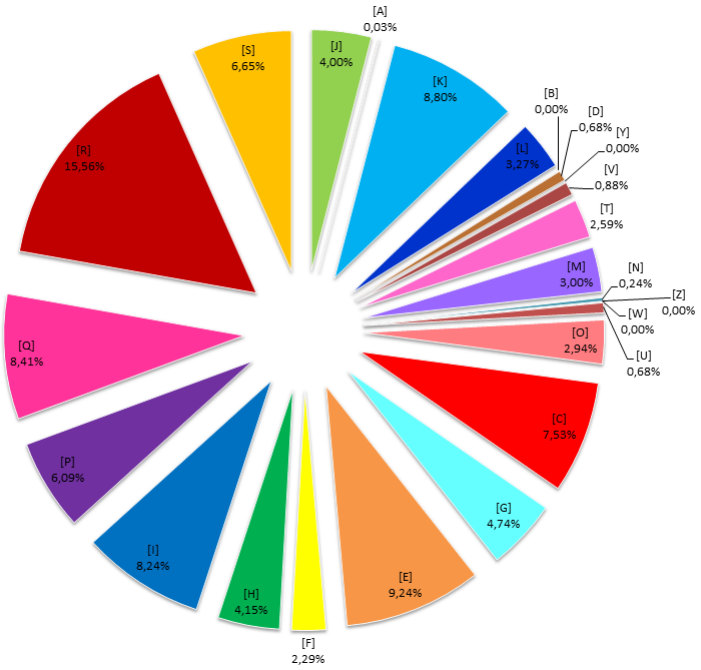



**Fig. 3.** Distribution of annotated ORFs of GO 06 isolate in COG categories. In legends, each letter inside square brackets corresponds to a COG category which is followed by the respective percentage of ORFs that could be classified in this category. [B] Chromatin structure and dynamics; [C] Energy production and conversion; [D] Cell cycle control, cell division, chromosome partitioning; [E] Amino acid transport and metabolism; [F] Nucleotide transport and metabolism; [G] Carbohydrate transport and metabolism; [H] Coenzyme transport and metabolism; [I] Lipid transport and metabolism; [J] Translation, ribosomal structure and biogenesis; [K] Transcription; [L] Replication, recombination and repair; [M] Cell wall/membrane/envelope biogenesis; [N] Cell motility; [O] Posttranslational modification, protein turnover, chaperones; [P] Inorganic ion transport and metabolism; [Q] Secondary metabolites biosynthesis, transport and catabolism; [R] General function prediction only; [S] Function unknown; [T] Signal transduction mechanisms; [U] Intracellular trafficking, secretion, and vesicular transport; [V] Defense mechanisms.

## 4 Conclusions and Perspectives

By constructing a computational pipeline composed of several bioinformatic tools, we managed to assemble *M. massiliense* strain GO 06's entire genome. It was also possible to annotate most of its conserved ORFs. The genome comparison with other mycobacteria revealed a higher similarity with pathogenic species, specially with *M. abscessus*, than non-pathogenic ones.

Further annotation will be performed regarding *M. massiliense*'s virulence factors and definition of its metabolic pathways. Comparative genomics will also

have an important role in this project's next steps, as to clarify *M. massiliense*'s classification and species relatedness.

## References

1. Adékambi, T., Reynaud-Gaubert, M., Greub, G., Gevaudan, M.J., Scola, B.L., Raoult, D., Drancourt, M.: Amoebal coculture of *Mycobacterium massiliense* sp. nov. from the sputum of a patient with hemoptoic pneumonia. Journal of Clinical Microbiology **42** (December 2004) 5493–5501
2. Petrini, B.: *Mycobacterium abscessus*: an emerging rapid-growing potential pathogen. APMIS **114** (May 2006) 319–328
3. Cardoso, A.M., Junqueira-Kipnis, A.P., Kipnis, A.: *In Vitro* antimicrobial susceptibility of *Mycobacterium massiliense* recovered from wound samples of patients submitted to arthroscopic and laparoscopic surgeries. Minimally Invasive Surgery **2011** (2011)
4. Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., Salzberg, S.L.: Versatile and open software for comparing large genomes. Genome Biology **5** (January 2004)
5. Warren, A.S., Setubal, J.C.: The genome reverse compiler: an explorative annotation tool. BMC Bioinformatics **10** (January 2009)

# *APÊNDICE B – Artigo apresentado no BSB 2013*

## A pipeline to characterize virulence factors in *Mycobacterium massiliense* genome

Guilherme Menegói[1], Tainá Raiol[1], João Victor de Araújo Oliveira[2], Edans Flávius de Oliveira Sandes[2], Alba Cristina Magalhães Alves de Melo[2], Andréa Queiroz Maranhão[1], Ildinete Silva-Pereira[1], Anamélia Lorenzetti Bocca[1], Ana Paula Junqueira-Kipnis[3], Maria Emília M. T. Walter[2], André Kipnis[3], and Marcelo de Macedo Brígido[1]

[1] Department of Cellular Biology, Institute of Biology, University of Brasilia, 70910-900, Brasília, DF, Brazil
[2] Department of Computer Science, Institute of Exact Sciences, University of Brasilia, 70910-900, Brasília, DF, Brazil
[3] Department of Microbiology, Immunology, Parasitology, and Pathology, Federal University of Goiás, 74605-050, Goiânia, GO, Brazil

**Abstract.** Virulence factors represent crucial molecular features for understanding pathogenic mechanisms. Here we describe a pipeline for *in silico* prediction of virulence factor genes in *Mycobacterium massiliense* genome that could be easily used in many other bacterial systems. Some few methods for this characterization are described in the literature, however these approaches are usually time-consuming and require information not always readily available. Using the proposed pipeline, the number and the accuracy of predicted ORF annotation were increased, and a broad identification of virulence factors could be achieved. Based on these results, we were able to construct a general pathogenic profile of *M. massiliense*. Furthermore, two important metabolic pathways, production of siderophores and bacterial secretion systems, both related to *M. massiliense*'s pathogenicity, were investigated.

**Keywords:** genome; rapid growing mycobacteria; pipeline; bioinformatics; nosocomial infections; virulence factors, metabolic pathways

## 1 Introduction

### 1.1 *Mycobacterium massiliense*

With the improvement of culture and identification techniques, the number of reported medical cases related to nontuberculous mycobacteria (NTM) has been greatly increased during the last few years [6]. Among those, mycobacteria of the *Mycobacterium chelonae-Mycobacterium abscessus* group composed of *M. chelonae*, *M. immunogenum* and, particularly, *M. abscessus*, arose as one of the most important opportunistic pathogens [14].

In 2004, Adékambi *et al.* [2] assigned a novel species for a closely related isolate, *Mycobacterium massiliense*, a representative species of rapidly growing

mycobacteria (RGM). RGM have important implications in human diseases, as they are frequently associated to infections among immunocompromised patients as well as wound, skin, and soft tissue infections [14]. Additionally, these bacteria are naturally resistant to several classes of antibiotics, particularly to antituberculousis drugs. *M. massiliense* is characterized as a strictly aerobic, non-spore-forming, nonmotile, acid-fast, gram-positive rod that shares 100% of its 16S rRNA sequence with *Mycobacterium abscessus*. However, there is still intense debate in the scientific community whether or not *M. massiliense* should be considered a new species or simply a *M. abscessus* strain.

Ever since its description, *M. massiliense* has been increasingly reported as the responsible for soft tissue infection outbreaks. At the Midwest Region of Brazil, a major infection outbreak has been recently reported along with the association to antibiotic and disinfectants resistances that may have contributed to the difficulty in controlling the spread of this strain. In a previous work, our group sequenced and characterized the genome of a *M. massiliense* strain, which was isolated from wound samples of patients submitted to arthroscopic and laparoscopic interventions in Goiânia, Brazil [7]. This strain has been since then identified as "GO 06" and its complete genome is already available in GenBank [15].

### 1.2 Virulence factors and their role in pathogenesis

Some bacteria are known to be extremely virulent pathogens with the ability to cause infectious diseases, e.g. tuberculosis or salmonellosis. Pathogenic bacteria must be able to enter its host, to survive and to replicate inside the host cell, while avoiding the mechanisms of host cell protection. Therefore, bacteria present a set of molecular features in order to bypass or overcome the host defenses, which are commonly called virulence factors. Here we discuss two major virulence factor systems, which seem to play a decisive role in *M. massiliense*'s pathogenicity: the siderophores and bacterial secretion systems production pathways.

Siderophores are ferric ion specific chelating agents whose main role is to scavenge iron from the environment and make it available to the microbial cell. It is well known that the siderophore system is correlated to the virulence of some organisms, like *Yersinia enterocolitica* and *Erwinia chrysanthemi* [13], and there are evidences it has a function in *M. massiliense*'s pathogenicity, as ferric iron is an essential macronutrient for bacterial growth.

The pathogenicity of some bacteria, however, depend on their ability to secrete virulence factors, which can be displayed on the bacterial cell surface, secreted into the extracellular medium, or directly injected into a host cell. In Gram-negative bacteria, six systems have been described with this function, the bacterial secretion systems I-VI. However, recent studies have provided evidence that in Gram-positive bacteria, such as *M. massiliense*, an alternative protein-secretion system exists, the type VII secretion system (T7SS), which has five copies through the genome, named ESX-1 to ESX-5 [1].

### 1.3 A pipeline for virulence factor analysis

In this article, we propose a pipeline for virulence factor identification and annotation, which was used in *Mycobacterium massiliense*'s genome. This pipeline could also be applied to other bacteria, possibly increasing the available data on bacterial pathogenesis and supporting the development of counter strategies against pathogenic microorganisms.

Few *in silico* methods have been described to predict virulence factors in bacteria, most of them based on machine learning strategies, which consider common molecular features of known virulence factors to predict new ones [4, 10]. These approaches rely on Support Vector Machine (SVM), a supervised learning strategy. For efficient characterization of virulence factors, these methods require a good quality input data, specially for the training phase, leading to time demanding programs. Furthermore, since virulence factors present a variety number of functions in the microbial cell, from a cell wall component to a secreted protein, finding common patterns to identify such molecules is really a challenging task.

There are also other methodologies based on phylogenetic information, which is difficult to get and therefore not always available [12]. In this context, we believe that our pipeline could be a simple yet very efficient alternative to characterize virulence factors in bacteria.

## 2 Methods

### 2.1 *M. massiliense* and *Bacillus anthracis* genomic data

*M. massiliense* GO 06 assembly data comprise a single chromosome, previously assembled by our group using MIRA [9], and two putative plasmids, named Plasmid I and Plasmid II, of roughly 60 and 96 kilobases, respectively. The ORFs (Open Reading Frames) were annotated with Genome Reverse Compiler (GRC), using a reference database composed of only mycobacteria genomes [16]. To validate the pipeline, all analysis were also done on the genome of a pathogenic bacteria from the Bacillus group, *Bacillus anthracis* str. Ames, whose sequences were downloaded from NCBI (AN:AE016879).

### 2.2 Pipeline proposal

An overview of the proposed pipeline is shown in Figure 1. The pipeline is divided in four major steps: (1) Improvement of ORF annotation; (2) Identification of virulence factors; (3) Analysis of virulence factor genes in metabolic pathways; and (4) Construction of metabolic pathway maps. Steps 1 and 2 are automated by a single Perl script, as seen in Figure 2, which shows the management of the pipeline execution. The details of this pipeline are described below. All analyses were made on a desktop with an AMD Phenom II B95 X4 processor and 4 GiB of RAM, running Ubuntu 12.04.

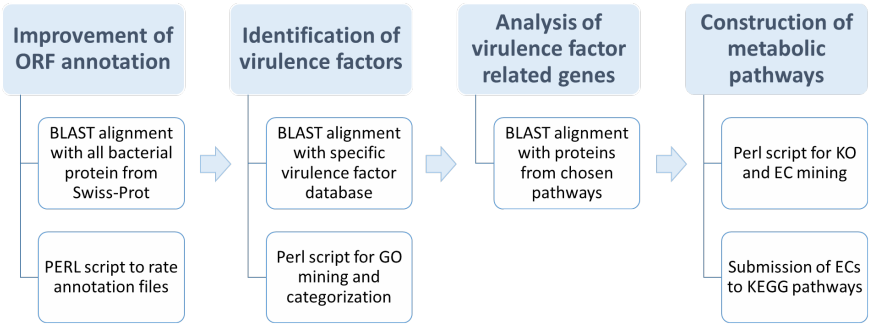



**Fig. 1.** Pipeline for identification and annotation of bacterial virulence factors.

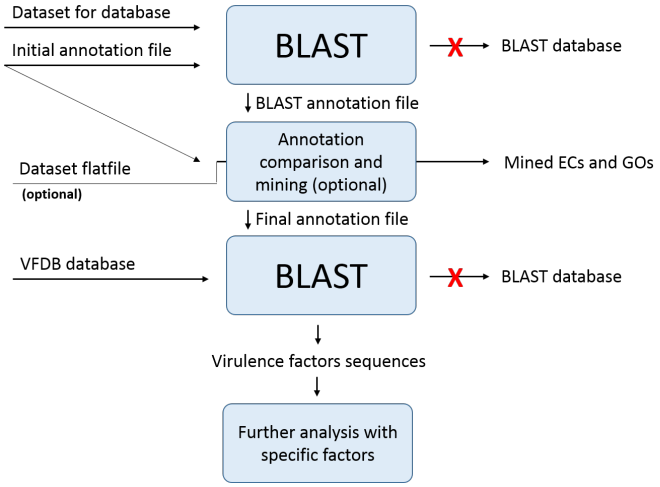



**Fig. 2.** Management of pipeline execution. The arrows leading to the boxes represent input files, while the arrows coming from them represent output files (those marked with a red 'x' are excluded by the end of the pipeline). All steps leading to the characterization of virulence factors are integrated in a single perl script.

**Improvement of ORF annotation** In order to have a well characterized sequence dataset, a BLAST [3] search with the initial annotation is performed, using a user-defined dataset for database construction. The E-value cut-off is also user-defined, having a default value of 1E-5.

The main part of the Perl script was designed to compare both the initial and the newly generated BLAST annotation files by assigning a score to each annotation and later choosing the highest score. For each BLAST query, the script verifies all the obtained hits, in order to choose the best annotation.

The presence of keywords such as "hypothetical" or "putative" penalizes the score of a hit, while complementary features (e.g., Gene Name (GN), Enzyme Commission number (EC)) increases its score. This scoring method was defined in order to penalize uncharacterized sequences and to favour well annotated proteins, and both weights are user-defined upon initiation of the script. Initial tests with this pipeline suggested that penalizing keywords by 2 and adding 0.5 for complementary features avoided most bad annotation lines, therefore those weights were chosen as defaults.

The final gene annotation is the one with the highest score, obtained either from the initial file or from any of the BLAST hits. If many hits were equally good, the script chooses the one from the BLAST annotation file. In case of two or more BLAST hits with the same score, the one with lowest e-value is preferred.

The initial annotation of *M. massiliense* was provided by GRC, while the initial annotation file of *B. anthracis* was downloaded from NCBI. For both organisms, the dataset used for the construction of the BLAST database was comprised of all curated bacterial proteins from UniProtKB/Swiss-Prot [5] (329,037 sequences, as of April 2013). Both the E-value cut-off and scoring weights were the scripts' default.

**Identification of virulence factor genes** The annotation file generated in the last step is then compared to the bacterial protein database from VFDB (Virulence Factor Database) [8], a specialized repository of bacterial virulence factors. In this step, we are interested in finding genes related to virulence according to sequence similarity. Therefore, all ORFs with hits coming from the VFDB entries are considered virulence factors and stored in a new annotation file. The sequences filtered by this criterion had their GO (Gene Onthology) classification determined in the previous step, on an optional script module for managing flatfiles.

**Analysis of virulence factor genes in metabolic pathways** Considering that pathogenesis-related genes are often present in mycobacteria, we decided to analyse two of the most relevant metabolic pathways involved in this genus' pathogenicity: production of siderophores and bacterial secretion system proteins.

The gene sequences composing the chosen virulence pathways were downloaded from KEGG (Kyoto Encyclopedia of Genes and Genomes) [11]. By aligning the selected ORF sequences annotated as virulence factors with KEGG sequences, we could identify the genes related to the pathways and assign them a KO (KEGG Onthology). A Perl script was used in KO and EC mining, taking this information directly from the flatfile.

**Construction of metabolic pathways** These results were used in the construction of metabolic maps for both pathways through KEGG, highlighting the genes present in *Mycobacterium massiliense* GO 06 genome. This analysis

could not be performed for protein-secretion system VII (T7SS), since it is not yet fully described and there are no available corresponding pathway in KEGG.

## 3 Results and Discussion

The proposed pipeline allowed us to increase the number and accuracy of annotated ORFs. The assembled chromosome, which had initially 3,053 annotated ORFs, when compared to the final 3,388, showed an increase of 11% in ORF annotation. This also happened for both putative plasmids: Plasmid I initially had 33 annotated ORFs, which increased to 47 after being processed by this pipeline (42.4% increase), and Plasmid II, from 14 to 58 (314.3% increase) annotated ORFs. The data regarding the execution of the pipeline for both genomes can be found on Table 1.

After identification and selection of all the virulence factor related genes, they were classified into a few selected GO categories for an easier overview of *M. massiliense*'s virulence profile. Table 2 shows this distribution for *M. massiliense*'s chromosome and both putative plasmids. Out of the 807 genes found to be related to virulence, 387 (48%) were genes involved in the organism's metabolism, most of them related to lipids (32%). In addition, 139 genes (17.8%) could not be properly characterized since their functions were poorly annotated (either they had no assigned GO or it was too unspecific, like "ATPase"). The high number of genes that could not be assigned to any class shows that there is still a limitation in protein databases related to virulence factors, even though most of the annotation came from curated sources.

**Table 1.** General data regarding the script execution. Vague annotations, such as "uncharacterized protein" were considered incomplete annotations. "I" means initial annotation; "W" weighted; and "U" unweighted.

| | *M. massiliense* | | | *B. anthracis* | | |
|---|---|---|---|---|---|---|
| | I | W | U | I | W | U |
| Incomplete annotations | 1.659 | 997 | 997 | 2.503 | 1.923 | 2.088 |
| Improved annotations | - | 2.816 | 2.816 | - | 3.646 | 3.572 |
| Elapsed time | - | 1h28m19s | 1h28m55s | - | 2h26m31s | 2h25m15s |

Figures 3 and 4 show the number of detected genes in *M. massiliense* genome that compose the two studied metabolic pathways. Figures 6 and 7 depict the constructed metabolic pathway maps. Because of a severe limitation of KEGG pathways tool and database, some gene families were not correctly displayed and were highlighted manually, with an image editing software. This is the case of *mbt* genes, for example, which share the same EC number and could not be correctly displayed.

**Table 2.** Distribution of virulence factor genes from *M. massiliense* GO 06 chromosome and two putative plasmids into GO categories.

| GO Category | Chromosome | Plasmid I | Plasmid II |
|---|---|---|---|
| Translation, ribosomal structure and biogenesis | 5 | 0 | 0 |
| Transcription | 58 | 0 | 0 |
| Replication, recombination and repair | 5 | 8 | 0 |
| Cell cycle control, cell division, chromosome partitioning | 21 | 0 | 1 |
| Pathogenesis | 79 | 1 | 6 |
| Signal transduction mechanisms | 22 | 0 | 0 |
| Cell wall/membrane/envelope biogenesis | 9 | 0 | 0 |
| Stress response | 55 | 1 | 1 |
| Posttranslational modification, protein turnover, chaperones | 27 | 0 | 0 |
| Energy production and conversion | 14 | 0 | 0 |
| Carbohydrate transport and metabolism | 37 | 0 | 0 |
| Amino acid transport and metabolism | 49 | 0 | 0 |
| Nucleotide transport and metabolism | 7 | 0 | 0 |
| Cofactor transport and metabolism | 19 | 0 | 0 |
| Lipid transport and metabolism | 124 | 0 | 0 |
| Inorganic ion transport and metabolism | 46 | 0 | 1 |
| Secondary metabolites biosynthesis, transport and catabolism | 66 | 0 | 0 |
| Antibiotic biosynthesis | 25 | 0 | 0 |
| General function prediction only | 139 | 0 | 0 |

The majority of siderophore genes are from the *entA* family (EC 1.3.1.28, 81 copies), which might be explained by the fact that it is involved in the production of 2,3-Dihydroxybenzoate, a necessary precursor of the siderophores vibriobactin, enteroxelin, bacillibactin and myxochelin (Figures 3 and 6). This result indicates that the precursor could possibly be needed in a higher quantity in the bacterial cell. Other gene families involved in the production of siderophore precursors, such as *menF* (EC 5.4.4.2), *mbtI* (EC 4.2.99.21) and *entB* (EC 3.3.2.1), however, present a much lower number of genes in *M. massiliense* GO 06's genome. While most siderophore molecules seem to be produced by *M. massiliense*, yersiniabactin, vibriobactin and pyochelin are either not produced or have an alternative structure, as evidenced by the presence of only part of the gene families coding the precursor chemical structures.

Genes from all secretion systems could be identified, most of them related to Type I (49 genes), followed by genes from Type III (27 genes). Additionally, 15 genes related to T7SS were found in *M. massiliense*'s genome, as shown in Figure 5, the majority of which belonged to the *EccA* gene family, whose products have ATPase activity, probably supplying energy for this system's functionality. However, none of the pathways were completely characterized, and key elements for the functionality of each system are still missing.

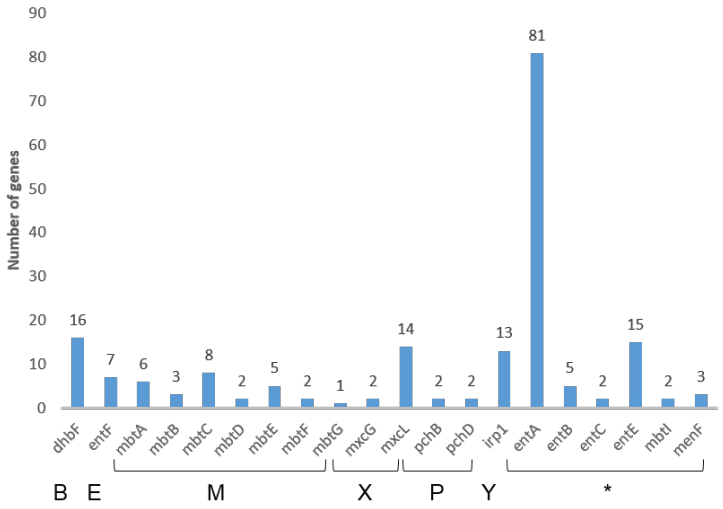



**Fig. 3.** Distribution of *M. massiliense* GO 06 genes related to siderophore production. The y axis represents the number of genes from a given family present in *M. Massiliense*'s genome. A letter code was assigned to each gene family, representing the siderophore production in which they are involved: Bacillibactin (B), Enterobactin (E), Mycobactin (M), Myxochelin (X), Pyochelin (P) and Yersiniabactin (Y). Gene families marked with '*' are either involved in the production of siderophore precursors or in more than one of the previously cited molecules.

It is noteworthy that the characterization of virulence factors has never been done before for *M. massiliense* and the results obtained in the present study are valuable, mainly considering the relevance of this organism as a pathogen and as a study model to *M. tuberculosis*. By our sequence similarity approach, the greater the amount of information in virulence factor databases, the better the final quality of our characterization. This might have affected our analysis, specially for T7SS, a recently characterized system that does not have many related sequences available in databases. Therefore, additional data, such as transcriptome sequences, could be useful to improve this characterization.

## 4 Conclusion

In this work, we propose a pipeline to identify virulence factors, which can also be used to improve annotation in bacteria. Other approaches exist to solve this problem, e.g., methods based on Support Vector Machine, which are time demanding, and those based on phylogenetic information that are not always available. We believe that our pipeline is easy to implement and fast to produce refined results, when compared to the other tools. Even though we applied this pipeline to identify *M. massiliense*'s genes, it could be easily used in the characterization of other pathogenic bacteria, regardless their classification.

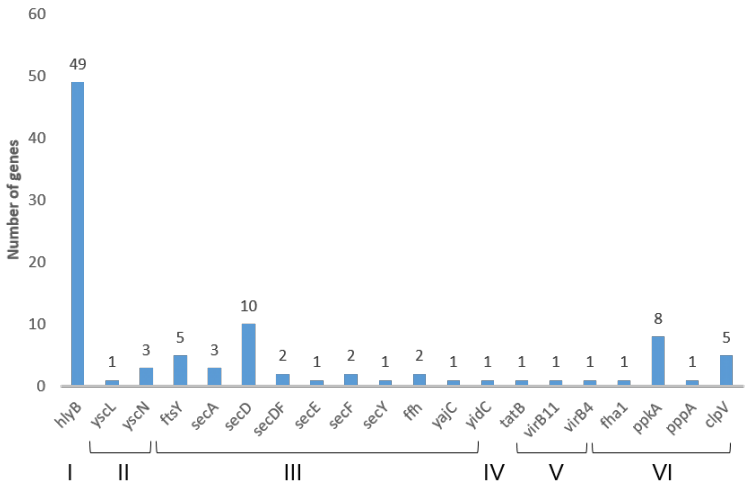



**Fig. 4.** Distribution of *M. massiliense* GO 06 genes related to bacterial secretion systems production. The y axis represents the number of genes from a given family present in *M. massiliense*'s genome. A number was assigned to each gene family, representing the secretion system production in which they are involved (I to VI).

By analyzing *M. massiliense*'s virulence factors, we could define an overview of this organism's pathogenicity profile and verify the existence of important mycobacterial genes, which validate the consistency of the assembled genome. In addition, *in silico* data regarding isolate GO 06's genes related to siderophores and bacterial secretion system proteins were obtained, which could prove useful for developing strategies to control *M. massiliense* related outbreaks.

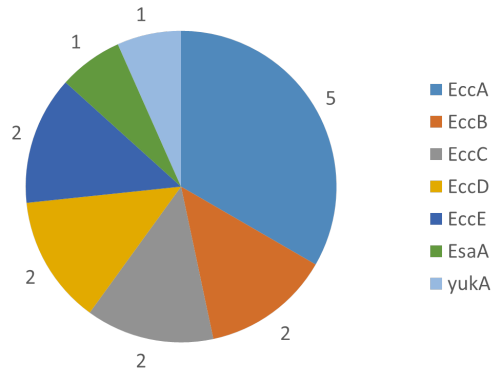



**Fig. 5.** Distribution of *M. massiliense* GO 06 genes related the Type VII secretion system production in *M. massiliense*'s genome.

**Acknowledgements.** This work was supported by CNPq (grant numbers 301198/2009-8 and 564243/ 2010-8). G.M.R and M.E.M.T.W. were supported by research fellowships from CNPq, and T.R. by research fellowship from CAPES.

## References

1. Abdallah, A., Gey van Pittius, N., Champion, P., Cox, J., Luirink, J., Vandenbroucke-Grauls, C., Appelmelk, B., Bitter, W.: Type "vii" secretion - mycobacteria show the way. Nature Reviews Microbiology 11, 883–891 (2007)
2. Adékambi, T., Reynaud-Gaubert, M., Greub, G., Gevaudan, M., La Scola, B., Raoult, D., MicGilbert: Amoebal coculture of *Mycobacterium massiliense* sp. nov. from the sputum of a patient with hemoptoic pneumonia. Journal of Clinical Microbiology 42, 5493–5501 (2004)
3. Altschul, S., Gish, W., Miller, W., Myers, E., Lipman, D.: Basic local alignment search tool. Journal of Molecular Biology 215(3), 403–410 (1990)
4. Andreatta, M., Nielsen, M., Aarestrup, F., O., L.: *In silico* prediction of human pathogenicity in the $\gamma$-proteobacteria. Plos One 5, e13680 (2010)
5. Bairoch, A., Boeckmann, B.: The "swiss-prot" protein sequence data bank. Nucleic Acids Research 20, 2019 (1992)
6. Carbonne, A., Brossier, F., Arnaud, I., Bougmiza, I., Caumes, E., Meningaud, J.P., Dubrou, S., Jarlier, V., Cambau, E., Astagneau, P.: Outbreak of nontuberculous mycobacterial subcutaneous infections related to multiple mesotherapy injections. Journal of clinical microbiology 47(6), 1961–1964 (2009)
7. Cardoso, A., Junqueira-Kipnis, A., Kipnis, A.: *In vitro* antimicrobial susceptibility of *Mycobacterium massiliense* recovered from wound samples of patients submitted to arthroscopic and laparoscopic surgeries. Minimally Invasive Surgery 2011, 1–4 (2011)
8. Chen, L., Yang, J., Yu, J., Yao, Z., Sun, L., Shen, Y., Jin, Q.: "vfdb": a reference database for bacterial virulence factors. Nucleic Acids Research 33, D325–328 (2005)
9. Chevreux, B., Wetter, T., Suhai, S.: Genome sequence assembly using trace signals and additional sequence information. Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB) 99, 45–56 (1999)
10. Garg, A., Gupta, D.: VirulentPred: a SVM based prediction method for virulent proteins in bacterial pathogens. BMC Bioinformatics 9, 62 (2008)
11. Kanehisa, M., Goto, S.: KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research 28, 27–30 (2000)
12. Nanni, L., Lumini, A.: An ensemble of support vector machines for predicting virulent proteins. Expert Systems with Applications 36(4), 7458–7462 (2009)
13. Neilands, J.: Siderophores: Structure and function of microbial iron transport compounds. The Journal of Biological Chemistry 270, 26723–26726 (1995)
14. Petrini, B.: *Mycobacterium abscessus*: an emerging rapid-growing potential pathogen. APMIS 114, 319–328 (2006)
15. Raiol, T., Ribeiro, G., Maranhão, A., Bocca, A., Silva-Pereira, I., Junqueira-Kipnis, A., Brgido, M., Kipnis, A.: Complete genome sequence of *Mycobacterium massiliense*. Journal of Bacteriology 194, 5455 (2012)
16. Warren, A., Setubal, J.: The genome reverse compiler: an explorative annotation tool. BMC Bioinformatics 10, 35 (2009)

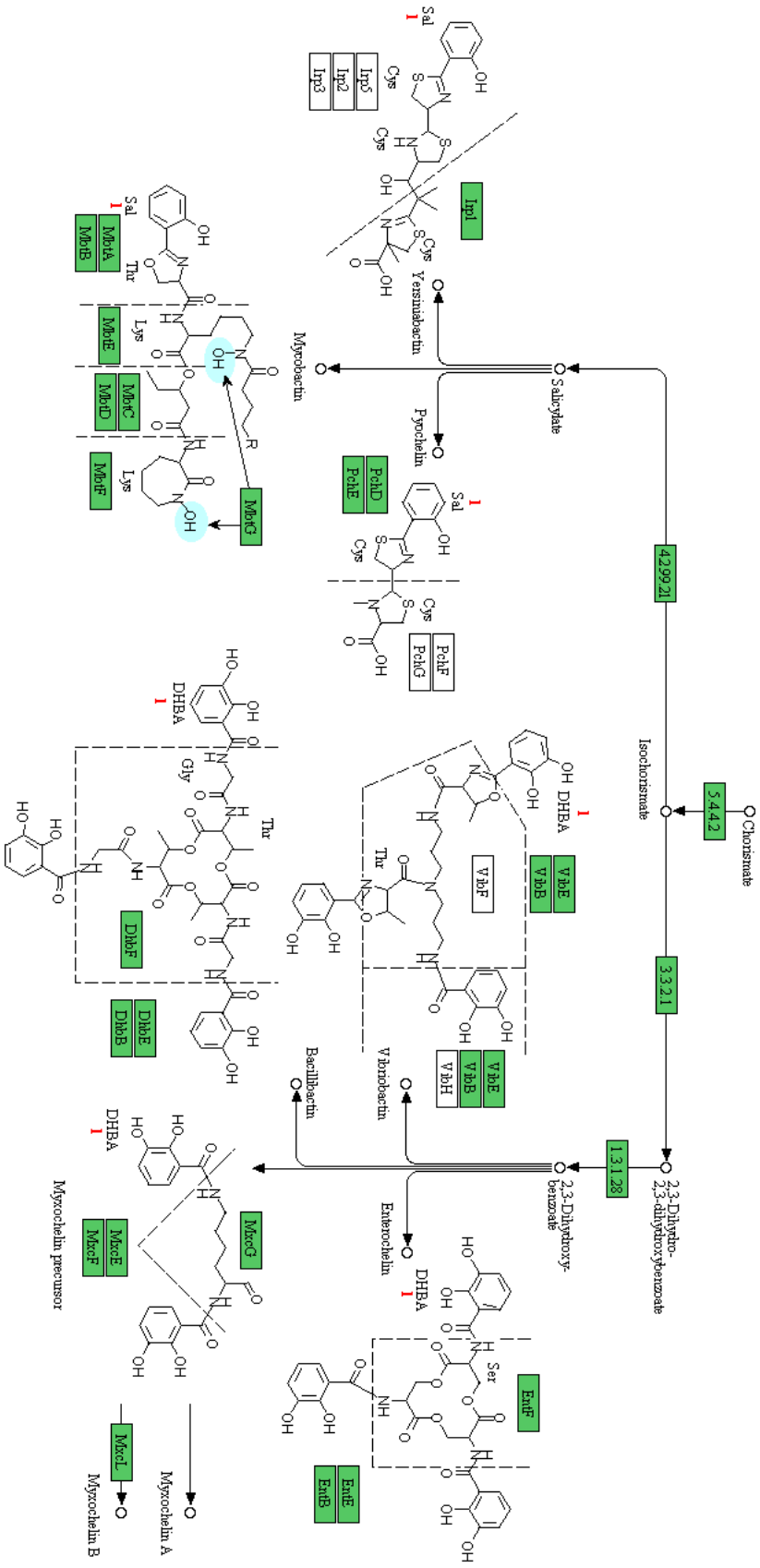



**Fig. 6.** Metabolic pathway map of siderophore production. Gene families in *M. massiliense* GO 06's genome are marked.

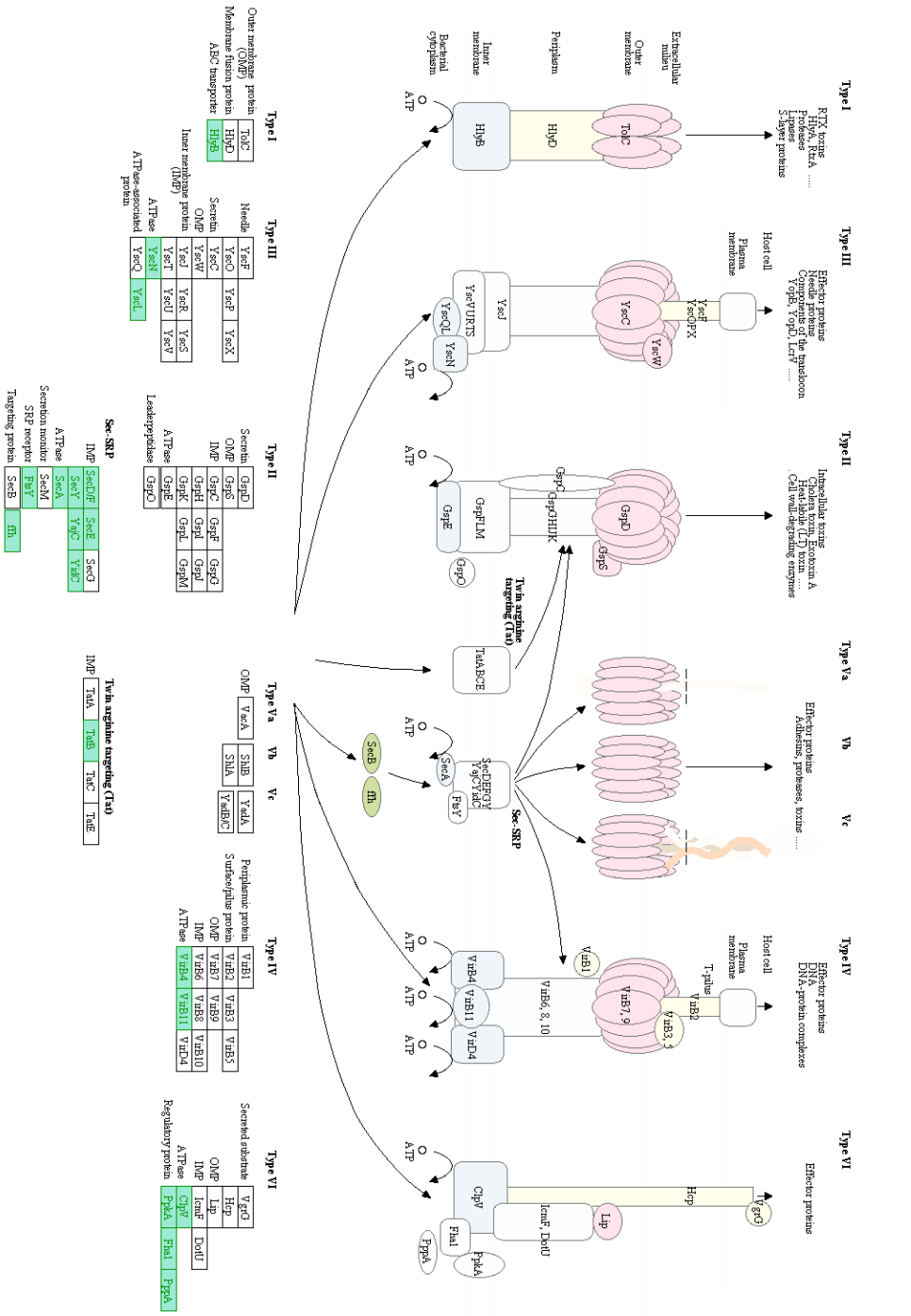



**Fig. 7.** Metabolic pathway map of bacterial secretion system productions (I-VI). Gene families in *M. massiliense* GO 06's genome are marked.