ANÁLISE COMPUTACIONAL DOS GENES ESSENCIAIS EM PROCARIOTOS: UMA ABORDAGEM COMPARATIVA DA FUNÇÃO, CONSERVAÇÃO E ORGANIZAÇÃO GENÔMICA

ANA LAURA GRAZZIOTIN

UNIVERSIDADE ESTADUAL DO NORTE FLUMINENSE DARCY RIBEIRO
Campos dos Goytacazes
2015

ANÁLISE COMPUTACIONAL DOS GENES ESSENCIAIS EM PROCARIOTOS: UMA ABORDAGEM COMPARATIVA DA FUNÇÃO, CONSERVAÇÃO E ORGANIZAÇÃO GENÔMICA

ANA LAURA GRAZZIOTIN

Tese apresentada ao Programa de Pós-graduação em Biociências e Biotecnologia, Centro de Biociências e Biotecnologia, Universidade Estadual do Norte Fluminense Darcy Ribeiro, como requisito parcial à obtenção do título de Doutora em Biociências e Biotecnologia.

Orientador: Prof. Dr. Thiago Motta Venancio

Campos dos Goytacazes
2015

ANÁLISE COMPUTACIONAL DOS GENES ESSENCIAIS EM PROCARIOTOS: UMA ABORDAGEM COMPARATIVA DA FUNÇÃO, CONSERVAÇÃO E ORGANIZAÇÃO GENÔMICA

ANA LAURA GRAZZIOTIN

Tese apresentada ao Programa de Pós-graduação em Biociências e Biotecnologia, Centro de Biociências e Biotecnologia, Universidade Estadual do Norte Fluminense Darcy Ribeiro, como requisito parcial à obtenção do título de Doutora em Biociências e Biotecnologia

Aprovada em 04 de dezembro de 2015.

Comissão examinadora:

_____
Prof. Rodrigo Nunes da Fonseca, Doutor em Genética e Genômica Funcional
Núcleo em Ecologia e Desenvolvimento Sócio Ambiental de Macaé - UFRJ

_____
Profª. Olga Lima Tavares Machado, Doutora em Bioquímica
Centro de Biociências e Biotecnologia – UENF

_____
Prof. Fabio Lopes Olivares, Doutor em Agronomia
Centro de Biociências e Biotecnologia - UENF

_____
Prof. Thiago Motta Venancio, Doutor em Bioinformática (Orientador)
Centro de Biociências e Biotecnologia - UENF

Ao meu esposo, colega de trabalho e mentor,

Newton de Medeiros Vidal.

Dedico.

AGRADECIMENTOS

A minha amada família pelo apoio emocional.

Aos que contribuíram nos trabalhos resultantes
desses quatro anos de doutorado.

À UENF/FAPERJ pela bolsa concedida de 2012 a 2014.

Ao National Institutes of Health e ao National Center for Biotechnology Information
pela oportunidade de estágio voluntário de 2014 a 2016.

SUMÁRIO

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CDS	*Coding sequence*
COG	*Cluster of orthologous groups*
DEG	*Database of essential genes*
DNA	Ácido desoxiribonucléico
Ka	Taxa de mutação não sinônima
Ks	Taxa de mutação sinônima
kb	Kilobases
MCA	*Multiple Correspondence Analysis*
NOG	*Nonsupervised orthologous groups*
OGEEdb	*Online gene essentiality database*
P	Valor de *P*
pb	Pares de bases
RNA	Ácido ribonucléico
°C	Graus Celsius
%	Porcentagem
3’	Extremidade três linha
5’	Extremidade cinco linha

RESUMO

A identificação de genes essenciais é crítica para o entendimento da fisiologia das espécies, para a proposta de novos alvos gênicos para drogas e para a descoberta do número mínimo de genes necessários para a sobrevivência de um organismo. Embora os grupos de genes essenciais de vários organismos tenham sido determinados por meio de técnicas de mutação em larga escala, estudos sistemáticos a respeito da conservação, do contexto genômico e das funções dos genes essenciais permanecem escassos. Neste trabalho, foram integrados dados de 17 estudos de genes essenciais identificados por meio de triagens cromossômicas *in vitro*, além de três estudos *in vivo*, compreendendo 15 espécies bacterianas e uma espécie de arqueia. Nossa abordagem comparativa dos dados desses 16 organismos revelou que os genes essenciais são, na maioria, provenientes de famílias monogênicas e tendem a ser mais conservados entre as espécies que os genes não essenciais. Em contrapartida, genes requeridos *in vivo* mostraram-se menos conservados quando comparados aqueles essenciais *in vitro* para a mesma espécie, sugerindo que estratégias distintas são empregadas quando o organismo está sob estresse imposto pelo sistema imune hospedeiro ou pelo ambiente nutricionalmente instável. Foram identificadas vias metabólicas análogas que provavelmente não seriam detectadas por meio de estratégias de predição de essencialidade baseadas em ortologia. Em *Streptococcus sanguinis*, por exemplo, genes de biossíntese de isoprenóides identificados como essenciais foram provavelmente transferidos horizontalmente a partir de arqueias. Este trabalho identificou um grupo de genes essenciais específicos de *Mycobacterium tuberculosis* e *Burkholderia pseudomallei* que poderiam representar possíveis alvos para drogas. Em termos de organização genômica, genes essenciais foram vistos como predominantemente localizados em operons, preferencialmente ocupando a primeira posição no sentido transcricional. Possivelmente, estes genes essenciais juntamente com suas regiões regulatórias, influenciem a transcrição do operon como um todo. Por fim, as características de conservação, organização e função dos grupos de genes essenciais identificadas neste trabalho são compartilhadas entre genomas de bactéria e arqueia. Por estarem presentes em grupos de organismos filogeneticamente distantes, estas caraterísticas podem também ter estado presentes no último ancestral comum destas espécies. Por outro lado, não podemos descartar que estas grandes tendências de arquitetura genômica em bactérias e arqueias tenham surgido independentemente.

Palavras-chave: genes essenciais, genômica comparativa, operons, procariotos, transposons.

ABSTRACT

Identification of essential genes is critical to understanding the physiology of a species, proposing novel drug targets and uncovering minimal gene sets required for life. Although essential gene sets of several organisms have been determined using large-scale mutagenesis techniques, systematic studies addressing their conservation, genomic context and functions remain scant. Here we have integrated 17 essential gene sets from genome-wide *in vitro* screenings and three gene collections required for growth *in vivo*, encompassing 15 Bacteria and one Archaea. We have refined and generalized important proposed theories using *Escherichia coli*. Essential genes were seen as typically monogenic and more conserved than nonessential genes. Genes required *in vivo* were seen less conserved than those essential *in vitro*, suggesting that more divergent strategies are deployed when the organism is stressed by host immune system and unstable nutrient availability. We have identified essential analogous pathways that would be probably missed by orthology-based essentiality prediction strategies. For example, *Streptococcus sanguinis* carries horizontally transferred isoprenoid biosynthesis genes that are wide-spread in Archaea. Genes specifically essential in *Mycobacterium tuberculosis* and *Burkholderia pseudomallei* were reported as potential drug targets. Moreover, essential genes were not only preferentially located in operons, but also occupying the first position therein, supporting the influence of their regulatory regions in driving transcription of whole operons. Finally, these important genomic features were shared between Bacteria and at least one Archaea, suggesting that high order properties of gene essentiality and genome architecture were probably present in the last universal common ancestor or evolved independently in the prokaryotic domains.

Keywords: essential genes, genome evolution, genome organization, operons, prokaryotes, transposon mutagenesis.

PREFÁCIO

Passados 20 anos dos primeiros estudos que tentaram encontrar o número mínimo de genes necessário para sustentar a vida celular, os cientistas ainda trabalham na obtenção de uma célula sintética auto-replicativa eficiente. Uma célula inteiramente artificial ainda não é disponível para produção em larga escala, entretanto, algumas tentativas de construir a célula sintética têm sido produtivas em ambiente de laboratório [1-3]. Partindo do nível mais básico para sustentar a vida, algumas características devem ser levadas em conta ao tentar construir uma célula:

- Qual a finalidade da célula a ser construída?
- Em qual ambiente a célula estará adaptada para sobreviver?
- Quais funções a célula irá necessitar para sobreviver e manifestar o fenótipo desejado? Sabendo disso, pode-se planejar o grupo de genes necessários para construir o genoma.
- Conhecer a organização e a regulação dos genes mantidos na célula para permitir sua sobrevivência e a expressão do fenótipo desejado.

Embora estes conceitos pareçam lógicos atualmente, o conhecimento a respeito da obtenção de uma célula sintética foi construído gradualmente ao longo dos anos, desde as primeiras discussões sobre o genoma mínimo e sobre a identificação de genes essenciais.

Inicialmente, tanto análises computacionais quanto experimentais foram empregadas para estabelecer o grupo mínimo de genes que poderia sustentar a vida procariótica. As análises computacionais são baseadas nas análises comparativas dos genomas procarióticos por meio da pesquisa de genes conservados entre os organismos estudados. As análises experimentais são baseadas nas tentativas de redução do genoma de um determinado organismo por meio de deleções de genes ou fragmentos cromossômicos. As duas abordagens, computacional e experimental, são complementares e foram fundamentais para o desenvolvimento de conceitos importantes acerca do genoma mínimo. Além do entendimento sobre quais genes ou regiões do genoma são indispensáveis para a célula, também se faz necessário entender como estas estruturas se organizam na célula. Por meio da combinação de estudos computacionais e experimentais, tem-se melhorado a compreensão sobre a organização (conservação dos genes e operons, sintenia gênica, uso

de códons e arquiteturas de domínios protéicos) e complexidade funcional (nível de expressão, dispensabilidade, interatoma, redes regulatórias) dos organismos, o que nos possibilitou redefinir o conceito e as perspectivas relacionadas à célula mínima ao longo das últimas duas décadas.

Neste estudo de tese é abordada a história da busca pelo genoma mínimo, envolvendo discussões sobre as mudanças na definição de célula mínima com a disponibilidade de comparar uma diversidade de organismos e sobre os avanços nos estudos de essencialidade com a evolução e disponibilidade de novas tecnologias. Como contribuição original desta tese destacam-se as análises comparativas realizadas para os dados de essencialidade de diversos organismos procarióticos publicamente disponíveis até março de 2013. Descrevo sobre as tendências encontradas para todos os organismos estudados nos aspectos de funcionalidade, conservação e organização dos genes essenciais. Da mesma forma, descrevo as características únicas de essencialidade identificadas em alguns organismos e como o conhecimento destas funções particulares e críticas para organismos patogênicos poderia nos ajudar a combater ou evitar infecções hospitalares ou na comunidade.

Este trabalho de tese encontra-se publicado no periódico FEBS Journal no ano de 2015 (doi: 10.1111/febs.13350 [4]). Na seção de resultados e discussão desta tese, algumas tabelas de resultados contendo um número extenso de linhas e colunas serão citadas como Tabela suplementar, material online. Devido ao tamanho dessas tabelas, seria inviável anexá-las ao documento de tese e assim, sugiro a consulta do material no site da revista (http://onlinelibrary.wiley.com/doi/10.1111/febs.13350/suppinfo) em que foram publicadas [4].

REVISÃO BIBLIOGRÁFICA

Os dois domínios procarióticos, Bacteria e Archaea, são bastante distintos [5] e adaptados a uma ampla variedade de nichos [6]. O tempo de divergência entre os maiores grupos de Bacteria ocorreu entre 2,5 a 3,2 bilhões de anos enquanto entre os maiores grupos de Archaea, ocorreu entre 3,1 e 4,1 bilhões de anos [7]. Os genomas procarióticos foram e têm sido moldados por meio de vários mecanismos, dentre eles, as pressões de seleção, o tamanho da população, a ocorrência de *bottlenecks* populacionais, eventos de mutação e recombinação, presença de elementos genéticos móveis, dentre outros [8], resultando em grande variabilidade em relação ao tamanho dos genomas e seus respectivos conteúdos gênicos entre diferentes grupos filogenéticos [9,10]. Entretanto, mesmo com os organismos conhecidos sendo tão distintos morfológica e fisiologicamente, um grupo de funções básicas é compartilhado entre os organismos existentes. Essa já era uma suspeita mesmo antes da aplicação das tecnologias de sequenciamento de DNA ao procariotos e eucariotos, acreditando-se que um grupo de genes seria comum entre esses organismos apesar da distância evolutiva.

1. O genoma mínimo baseado em análises computacionais

A ideia de um grupo de genes compartilhado entre os organismos foi fortalecida com o sequenciamento dos dois primeiros genomas bacterianos. A bactéria *Haemophilus influenzae* Rd linhagem KW20 foi o primeiro organismo cultivável em meio desprovido de células (alguns vírus haviam sido sequenciados previamente) a possuir seu genoma sequenciado [11]. Em seguida, *Mycoplasma genitalium* foi a bactéria escolhida para o projeto de sequenciamento do genoma [12]. Imediatamente após a publicação destes genomas, o primeiro trabalho de genômica comparativa destes dois organismos, conduzido pelo grupo do Dr. Koonin, propôs o conceito de genoma mínimo [13]. As bactérias estudadas, *H. influenzae* e *M. genitalium*, divergiram a partir de seu último ancestral comum há aproximadamente 1,5 bilhões de anos [13]. Ambos os organismos apresentam genomas considerados pequenos, com 1.743 genes codificadores de proteínas em *H. influenzae* e apenas 470 genes codificadores de proteínas em *M. genitalium*. Além disso, *H. influenza* é representante do

grupo das bactérias gram-negativas enquanto *M. genitalium* é um organismo relacionado às bactérias gram-positivas. Devido a essas características, acreditou-se que o grupo de genes conservados entre estas duas bactérias – 240 genes identificados na análise dos genes ortólogos – se aproximaria do grupo mínimo de genes essenciais para sustentar qualquer forma de vida [13]. Entretanto, os autores perceberam que muitas enzimas importantes para o metabolismo não estavam presentes no grupo de genes conservados, indicando que um genoma contendo apenas o grupo mínimo de genes apresentado no trabalho não seria capaz de sustentar uma célula viável. Aquelas funções para as quais os genes não estavam presentes no grupo de genes conservados possivelmente eram realizadas por genes não ortólogos que evoluíram para executar a mesma função em organismos distintos (Figura 1). Após uma cuidadosa inspeção manual em busca das funções indispensáveis para a célula, porém perdidas na análise de ortologia, foram identificados 22 casos de funções realizadas por genes não ortólogos [13]. Assim, o grupo de genes mínimo foi complementado, totalizando 256 genes (excluindo seis casos de redundância funcional) [13]. Por este motivo, logo no início dos estudos do genoma mínimo, tornou-se evidente que os pesquisadores deveriam procurar por um grupo mínimo de funções celulares ao invés de um grupo mínimo de genes.

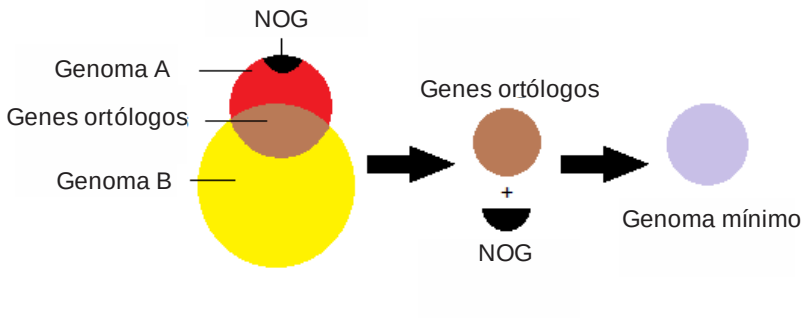


Figura 1 - Abordagem computacional para a determinação do genoma mínimo, levando em consideração os genes responsáveis por funções indispensáveis, porém não ortólogos entre os genomas estudados.
Fonte: ilustração adaptada a partir da referência [14].

A importância de um grupo de funções e dos genes não ortólogos foi fortalecida à medida que um maior número de genomas se tornou disponível para os estudos de genômica comparativa. Um desses trabalhos propôs um genoma mínimo composto de 206 genes para uma bactéria fermentativa, assumindo um meio nutricional rico para o crescimento bacteriano [15]. Este estudo foi baseado na comparação dos genomas de quatro

organismos endossimbiontes bacterianos, uma vez que estes apresentam os menores genomas procarióticos e assim poderiam reter os genes mais importantes para a sobrevivência, tais como aqueles relacionados às funções de manutenção celular. O grupo de genes obtidos nesta análise (180 genes) foi então comparado ao genoma do *M. genitalium* e às listas de genes essenciais disponíveis para quatro bactérias à época (*E. coli*, *Bacillus subtilis*, *M. genitalium* e *Staphylococcus aureus*). Por fim, as funções consideradas críticas, porém perdidas na análise, foram inspecionadas manualmente e adicionadas ao grupo mínimo de genes proposto (206 genes). Outro estudo sobre o genoma mínimo comparou os genomas de 21 organismos pertencendo aos três domínios da vida (Bacteria, Archaea e Eukarya) [14]. Um total de 253 genes foi estimado como o genoma mínimo. Dentre os 253 genes, somente 81 genes eram genes universais, 75 genes foram conservados em todas ou quase todas as bactérias estudadas e tinham representantes de Archaea e Eukarya e 97 genes foram distribuídos entre Bacteria, Archaea e Eykarya [14].

As diferenças encontradas para o genoma mínimo entre os trabalhos descritos foram tanto quantitativas (256, 206 e 253 genes) quanto qualitativas. O grupo de funções mínimas derivado dos dois trabalhos comparativos de bactérias [13,15], compreendeu:

- Maquinaria de replicação DNA e traducional quase completa;
- Sistema transcricional incompleto, com poucos fatores de transcrição;
- Sistema de reparo e recombinação do DNA rudimentar;
- Pequeno número de proteínas chaperonas e transportadores de metabólitos;
- Metabolismo energético baseado em glicólise e fosforilação;
- Biosíntese limitada de lipídeos e cofatores;
- Nenhuma enzima de biossíntese de aminoácidos e nucleotídeos.

Quando os domínios Archaea e Eukarya foram incorporados na análise do genoma mínimo, o grupo de funções encontradas foi bastante alterado [14,16,17]. O grupo mínimo continuou enriquecido em genes relacionados à tradução, biogênese e estrutura ribossomal, com a maioria dos genes universais (53/81) pertencendo a esta classe funcional [14]. Em contraste, as maquinarias de replicação do DNA dos eucariotos e das arqueias são similares, porém constituídas por genes não ortólogos em relação ao sistema bacteriano [16,17], permanecendo nesta classe apenas poucos genes universais [14]. De fato, funções essenciais desempenhadas por genes não ortólogos foram também identificadas no grupo

de proteínas da replicação e mesmo para o sistema traducional [14]. Somado a isso, poucos casos de genes universais foram presentes para as funções de reparo e recombinação do DNA e para chaperonas. Estas duas categorias, na verdade, foram enriquecidas em proteínas conservadas em quase todas as bactérias e em alguns representantes dos outros dois domínios, mostrando um padrão filogenético disperso. A maioria dos casos de genes não ortólogos foi descrita para vias metabólicas como transporte e metabolismo de nucleotídeos e produção de energia. Dessa forma, tornou-se evidente que os casos de funções realizadas por genes não ortólogos são muito comuns e, dessa forma, muitas proteínas ou vias bioquímicas alternativas poderiam constituir a célula mínima desde que preencham as funções requeridas para a manutenção e divisão celular sob condições ambientais específicas. Em outras palavras, as análises baseadas em ortologia fornecem um *core* de funções para o genoma mínimo. Este *core* deve ser complementado com as funções desempenhadas por genes não ortólogos, que podem oferecer muitos candidatos alternativos. Portanto, várias versões do genoma mínimo podem ser obtidas.

2. O genoma mínimo baseado em análises experimentais

A identificação de genes indispensáveis em bactérias tem despertado interesse de alguns grupos de pesquisa há décadas. Desde 1970, estudos de expressão gênica descreveram genes indispensáveis em *E. coli* baseados no fenótipo resultante – como múltiplas alterações celulares ou morte – do uso de deleções pontuais ou mutações sítio-dirigidas [18,19]. Entretanto, somente em 1995 foi conduzido um estudo pioneiro na identificação sistemática de essencialidade de *B. subtilis*, com o objetivo de estimar o genoma mínimo desta bactéria [20], mesmo sem a disponibilidade do genoma sequenciado desta espécie. As tentativas de redução do genoma, resultaram em seis fragmentos genômicos reduzidos (totalizando 562 kb) que impediram a formação de colônias e foram definidos como fragmentos essenciais para *B. subtilis* em meio rico a 37°C. Entretanto, a identificação de genes essenciais ainda não era possível devido à falta de informação genômica naquele tempo.

Em 1996, o genoma do *M. genitalium*, a menor bactéria conhecida capaz de crescer em condições de laboratório, foi sequenciado. A disponibilidade da informação gênica desta bactéria permitiu definir experimentalmente, em 1999, o grupo mínimo de genes necessários

para o crescimento [21]. A bactéria *M. genitalium* e a espécie relacionada, *M. pneumoniae*, foram submetidas a mutações sistemáticas por transposons. Após a avaliação do crescimento dos mutantes nas duas espécies, uma análise das populações mutantes mostrou que 351 genes ortólogos entre os organismos não apresentavam inserções e um grupo de 265 a 350 genes foi proposto como o genoma mínimo de *M. genitalium* [21]. Este grupo de genes mínimos encontrado experimentalmente foi maior que aqueles grupos mínimos (256, 253 e 206 genes) teoricamente propostos anteriormente [13-15]. Inesperadamente, 10 genes propostos como essenciais por meio de análises computacionais (tais como DNA helicase, subunidade alfa da DNA polimerase e a proteína ribossomal L28) [14] foram alvos de inserção de transposons e os mutantes foram viáveis experimentalmente [21]. Entretanto, a viabilidade destes mutantes poderia ser explicada pela análise de populações mutantes misturadas. Esta estratégia pode resultar na compensação da função perdida em um mutante, por outro mutante distinto presente na cultura [22].

Em 2006, a essencialidade gênica em *M. genitalium* foi re-analisada pelo mesmo grupo usando uma metodologia mais rigorosa, além de atingir a saturação das inserções nos mutantes analisados individualmente [22]. Cem genes foram mutados e 387 genes foram considerados essenciais [22]. Em contraste com o trabalho anterior, genes da replicação do DNA não foram identificados com inserções. Entretanto, genes envolvidos nos sistemas de reparo e recombinação do DNA tais como as helicases ruvA, ruvB, recA, foram alvos de transposons e os mutantes foram recuperados. Embora poucos genes de reparo e recombinação sejam universais e presentes em um dado genoma mínimo [14], sem tal maquinaria, a célula provavelmente não sobreviveria no ambiente. Este mesmo trabalho estimou 6% de redundância gênica em *M. genitalium* [22]. Nestes casos, um gene essencial poderia ser alvo de transposon e ainda resultar em um mutante viável devido à compensação da função por outro gene. Por isso, embora genes parálogos envolvidos no metabolismo de fosfato, glicerol e lipoproteínas tenham recebido inserções e sejam tecnicamente dispensáveis, estes foram considerados para o genoma mínimo de micoplasma [22].

Desde o estudo dos micoplasmas [21], um enorme número de projetos de identificação experimental de genes essenciais tem sido conduzido. Embora uma variedade de tecnologias tenha sido desenvolvida para este propósito, foi o advento das tecnologias de sequenciamento de segunda geração que revolucionou os estudos de genes essenciais, permitindo a execução de um grande número de procariotos. Inicialmente, estes estudos

experimentais eram focados na identificação do grupo mínimo de genes necessários para determinado organismo. Recentemente, os objetivos dos estudos se direcionaram para associações sistemáticas entre genótipo-fenótipo, descoberta de novos alvos para drogas e entendimento do metabolismo e patogenicidade do organismo. Por esses motivos, os estudos experimentais têm se dedicado à identificação de genes essenciais predominantemente em organismos modelo, de importância médica ou comercial (Anexo 1).

Uma ampla variedade de técnicas foi desenvolvida para a identificação de genes essenciais (Figura 2). Alguns dos métodos desenvolvidos rapidamente caíram em desuso devido às limitações técnicas e de interpretação dos resultados, por exemplos o método baseado em RNA antisenso utilizado para *S. aureus* em 2001 [23], o método baseado em inserção por transposon seguida pela amplificação por PCR de sequências cromossômicas sobrepostas em 5 Kb utilizado para *H. influenzae* em 2001 [24] e o método baseado em inserção por transposon seguida pela análise de hibridização diferencial por meio da técnica de microarranjo realizado para *Pseudomonas aeruginosa* em 2007 e *S. aureus* em 2009 [25]. Por outro lado, os métodos baseados em deleção única de genes são bastante confiáveis para as análises de essencialidade. Este método foi adotado para *B. subtilis* em 2003 [26], *E. coli* em 2006 [27], *Acinetobacter baylyi* em 2008 [28] e *Streptococcus sanguinis* em 2011 [29]. Entretanto, devido à complexidade de manipulação, de obtenção de mutantes para cada gene alvo e do alto custo, este método raramente é utilizado. Em contrapartida, as técnicas baseadas em inserção por transposon seguida de sequenciamento de segunda geração, particularmente com os intrumentos da Illumina (MiSeq ou HiSeq), tem sido amplamente adotadas devido à facilidade de manipulação, rapidez de obtenção dos resultados, possibilidade de trabalhar em larga-escala, sensibilidade para interpretação dos resultados e baixo custo. Em consequência, desde 2011 um grande número de trabalhos de essencialidade para uma diversidade de organismos tem sido publicado utilizando variações da técnica denominadas InSeq, TRADIS, TnSeq, entre outros (Anexo 1).

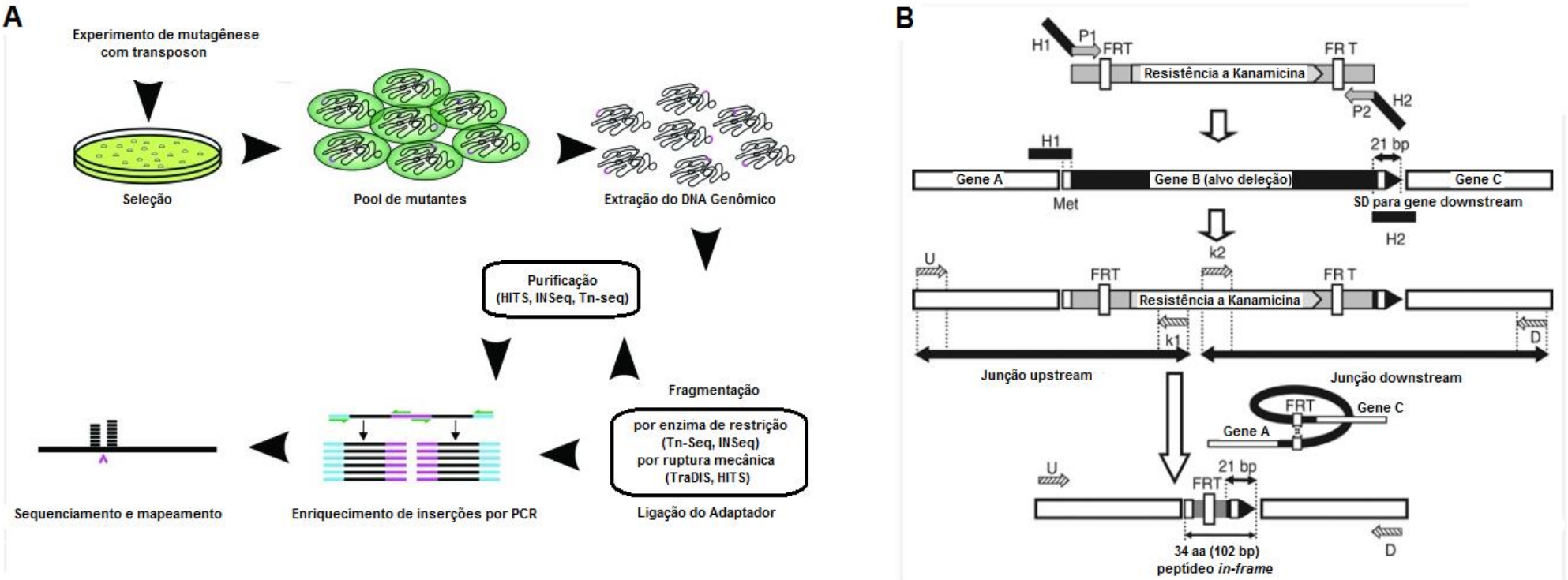


Figura 2 - As duas técnicas mais utilizadas para a identificação experimental de genes essenciais. A) representação esquemática genérica para a técnica baseada em mutação dos genes por transposons, seguida por sequenciamento de segunda geração (linhas rosas indicam os transposons), B) representação esquemática para a técnica de deleção gene a gene utilizada para Escherichia coli.
Fonte: adaptado a partir das referências [27,30].

Aproveitando o grande número de trabalhos de essencialidade, as listas de genes essenciais de procariotos e eucariotos foram compiladas em dois bancos de dados online (*Database of Essential Genes* ou DEG [31] e o *Online Gene Essentiality Database* ou OGEEdb [32]). O DEG tem sido o banco de dados mais utilizado devido à possibilidade da análise de similaridade no próprio banco entre as listas disponíveis no site ou destas com as listas de proteínas de interesse do usuário. A grande vantagem do DEG é a sua constante atualização com os novos dados experimentais gerados anualmente. Embora o banco de dados reúna experimentos sistemáticos, que consideram o genoma inteiro para a análise de essencialidade, não há discriminação quanto à saturação de inserções para os experimentos baseados em transposição. A saturação é um requisito importante relacionado à sensibilidade e a acurácia da técnica para definir os genes que são realmente essenciais. Por este motivo, para o desenvolvimento das análises apresentadas nesta tese, ao invés de usar os experimentos pré-compilados no DEG, foram cuidadosamente selecionados apenas os estudos no quais os autores dos trabalhos confirmaram a saturação ou a proximidade da saturação dos experimentos.

A maioria dos estudos apresenta uma lista de genes essenciais para o crescimento em meio rico, ou seja, em uma condição na qual todos os nutrientes necessários são fornecidos para a célula. Assim, a bactéria pode absorver os nutrientes do meio sem necessitar sintetizá-los *de novo*; aproveitando-se de uma condição sem estresse e sem competição, o grupo de genes essenciais pode ser reduzido a apenas aqueles genes da maquinaria celular básica, como genes da replicação do DNA e síntese de proteínas. De fato, vários estudos conduzidos independentemente em bactérias filogeneticamente distantes têm mostrado consenso na essencialidade de genes pertencentes às categorias funcionais relacionadas ao processamento da informação (replicação do DNA, transcrição e tradução). Por outro lado, os experimentos também mostram uma grande variabilidade no número de genes essenciais obtidos entre diferentes bactérias, assim como para a mesma bactéria em distintos experimentos [33]. A variabilidade dos genes essenciais não é apenas quantitativa, mas também qualitativa [34]. A versatilidade dos grupos de genes essenciais é influenciada por várias condições relacionadas às características de cada organismo e pelas características da técnica utilizada.

Dentre as características biológicas que influenciam a essencialidade de um organismo estão:

- Diferenças relacionadas ao estilo de vida do organismo;
- Herança genética distinta entre os grupos taxônomicos;
- Variação das vias metabólicas;
- Presença de genes isofuncionais;
- Diferenças fisiológicas.

Em contraste, os determinantes técnicos são (Anexo 1):

- Critério usado para a determinação do gene essencial;
- Condições de crescimento, o que inclui tipo e composição do meio (rico, mínimo, definido, caldo, sólido, *in vivo*, entre outros), temperatura, aerobiose ou anaerobiose, pressão, salinidade, entre outros;
- Abordagem técnica de identificação dos genes essenciais (deleção gene a gene, mutação por transposons e sequenciamento, mutação por transposon e análise por microarranjo, uso de RNA de interferência, entre outros).

As técnicas de mutação por transposon seguida por sequenciamento e as de deleção gênica têm sido as abordagens predominantes para os estudos e a variabilidade dos grupos de genes essenciais encontrados também é influenciada pelas limitações dessas técnicas. A maior limitação da técnica de deleção gene a gene está relacionada à taxa de transformação do organismo estudado. Uma taxa de eficiência de transformação baixa pode resultar na ausência de mutantes para genes não essenciais, que seriam então considerados como essenciais. Já a técnica de inserção por transposons é dependente de:

- Tipo do transposon usado. O transposon himar1 tem sido o mais utilizado. Este transposon apresenta inserção preferencial por sítios AT no genoma e assim, genes com poucos sítios AT poderiam ser erroneamente interpretados como essenciais [35];
- Complexidade da biblioteca (o número de mutantes obtidos carregando inserções únicas);
- Nível de saturação por transposons e cobertura do genoma. Sem a saturação de inserções no genoma, genes muito pequenos ou genes com poucos sítios de inserção podem não ser atingidos por transposons;
- Tempo de monitoramento do crescimento dos mutantes antes da análise. Mutantes para genes não essenciais de crescimento lento podem ser indetectáveis em um

período curto de cultivo ou render uma baixa contagem de *reads* para ser considerado não essencial;

- Método de sequenciamento. O sequenciamento de segunda-geração tem profundidade e ampla cobertura para identificar as junções gene-insertos de muitos mutantes simultaneamente e permitir resolução suficiente para identificar cada gene com inserto.

Muitos autores têm apresentado evidências sobre as influências das abordagens técnicas e das particularidades fisiológicas para cada experimento [21,22,27,36]. Um exemplo são os distintos grupos de genes encontrados para *M. genitalium* devido às diferenças na saturação de inserções entre dois experimentos [21,22], como mencionado acima. Dois outros experimentos independentes para *E. coli* K-12 apresentaram 620 [36] e 303 [27] genes essenciais e as suas discrepâncias foram relacionadas às distintas metodologias usadas e interpretação de essencialidade. Em contraste, devido às particularidades biológicas, os genes que são essenciais em um organismo podem ser não essenciais em outro, mesmo que de uma espécie relacionada ou entre linhagens da mesma espécie. Assim, os experimentos permitem identificar os genes essenciais que são específicos de um organismo. Um experimento realizado com duas linhagens de *Salmonella enterica*, cultivadas sob as mesmas condições e analisadas pela mesma técnica, revelou uma diferença de 85 genes requeridos entre *S. enterica* typhimurium SL1344 e *S. enterica* typhi Ty2 (Figura 3) [37]. A maior diferença entre os sorovares foi relacionada aos genes envolvidos na biossíntese da parede celular. Um grupo de quatro genes provavelmente presentes em um mesmo operon (SL0702, SL0703, SL0706 e SL0707) foi requerido apenas em *S. enterica* typhimurium, refletindo a adaptação evolutiva desses sorovares aos seus respectivos nichos [37]. Em contrapartida, genes relacionados à conversão de ferro III (operon fepBDGC) foram requeridos apenas em *S. enterica* typhi. Isto está de acordo com a utilização do ferro III no sangue por *S. enterica* typhi quando causando a infecção conhecida como febre tifoide. Por outro lado, a *S. enterica* typhimurium, que causa infecções intestinais, consegue utilizar o ferro a partir de várias fontes, incluindo o ferro II, solúvel no meio intestinal [37].

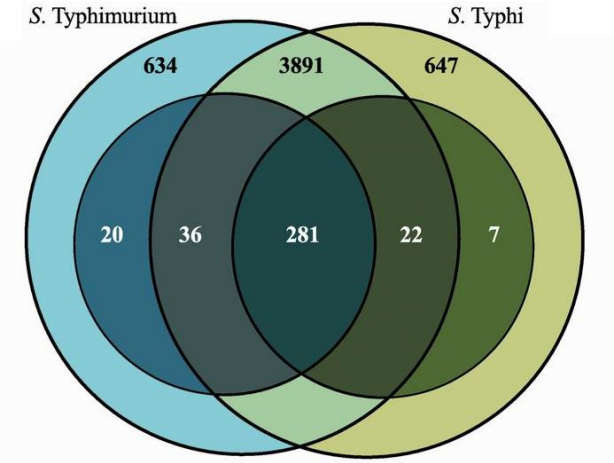


Figura 3 - Análise comparativa dos genes requeridos entre Salmonella enterica typhimurium e S. enterica typhi. Diagrama de Venn apresentando a análise de ortologia entre os genes codificadores das duas linhagens: números em preto e presentes nos círculos mais externos, indicam o total de genes codificadores de proteínas no genoma de cada linhagem e números em branco presentes nos círculos mais internos, indicam o total de genes requeridos nos experimentos. Números compreendidos entre os círculos representam genes compartilhados.
Fonte: ilustração adaptada a partir da referência [37].

Alguns experimentos de essencialidade *in vivo* também têm sido conduzidos. Esses estudos no hospedeiro são mais complexos que os estudos em meios de laboratório e por isso não são muitos os organismos com dados disponíveis. Entretanto, as análises *in vivo* são fundamentais porque elas refletem as condições mais aproximadas de situações reais. Por meio desses estudos podem-se definir quais são os requerimentos gênicos de um organismo durante o curso de infecção por uma bactéria patogênica, durante o processo de colonização por uma bactéria comensal ou de formação de biofilme por uma bactéria em um equipamento hospitalar ou na fibrose cística. Algumas bactérias alvo desses estudos tem sido o *Mycobacterium tuberculosis* durante a infecção em camundongo [38] e em macrófagos [39], *Francisella tularensis novicida* durante a infecção em camundongos [40,41], a *S. enterica* typhimurium durante infecção em camundongo [42], *Acinetobacter baumannii* em células do peritônio humano [43], *S. saguinis* durante a formação de biofilme [44], entre outros. O trabalho realizado para *Bacteroides tethaiotaomicron* [45] realizou simultaneamente as análises de genes essenciais *in vitro* e demonstrou características exclusivas de requerimento na comparação entre as duas condições. Enquanto 477 genes foram considerados essenciais em meio rico, apenas 280 genes foram requeridos dentre os mutantes recuperados do intestino hospedeiro. Os autores indicam uma sobreposição de 146 genes entre os genes requeridos a partir das duas condições, sendo que dentre os genes

requeridos exclusivos da coleção *in vivo*, estão presentes genes de diversas funções preditas, incluindo montagem de polissacarídeos e de estruturas baseadas em proteínas de superfície (genes BT1339-55, BT1953-7), genes relacionados à síntese de vitamina B12 (BT2090-1, BT2760) e complexos oxidoredutase (BT0616-22). Os experimentos *in vivo* fornecem uma oportunidade para responder questões relacionadas aos determinantes de patogenicidade e virulência de agentes patogênicos e quais os determinantes da persistência do indivíduo em uma condição estabelecida.

3. Os estudos comparativos de essencialidade entre organismos baseados na integração de análises experimentais e computacionais

3.1. Comparando a conservação dos genes essenciais

As taxas de evolução de sequência podem variar amplamente entre os genes codificadores de proteínas. Como a discriminação dos genes em essenciais e não essenciais está relacionada à adaptação e sobrevivência do organismo no meio, supostamente, genes essenciais apresentariam taxas de evolução reduzidas em decorrência da pressão de seleção purificadora. Em 2002, os dados de essencialidade de *Escherichia coli* foram analisados quanto à conservação de sequência nucleotídica por meio da razão entre a taxa de substituição não-sinônima (Ka) e a taxa de substituição sinônima (Ks) [46]. Este trabalho mostrou que os genes essenciais ($Ka/Ks = 8.40x10^{-2}$) são mais conservados que os genes não essenciais ($Ka/Ks = 11.67x10^{-2}$) em *E. coli* ($P < 6x10^{-6}$) [46]. Recentemente, outro trabalho reuniu os dados de essencialidade experimental de 23 bactérias e analisou a taxa de conservação de sequência destes organismos [47]. O estudo confirmou que os genes essenciais apresentam taxas de mutação menores quando comparados aos genes não essenciais na maioria das bactérias (Figura 4). Quando as taxas de mutação foram relacionadas às categorias funcionais, foi observado que os genes essenciais pertencentes às categorias de “Tradução, biogênese e estrutura do ribossomo” (J), “Transcrição” (K), “Replicação, recombinação e reparo” (L), “Metabolismo e transporte de carboidratos” (G), “Metabolismo e transporte de coenzimas” (H) e “Metabolismo e transporte de lipídeos” (I) foram significativamente ($P < 0.01$) mais conservados que os genes não essenciais nas mesmas categorias em mais da metade dos organismos estudados.

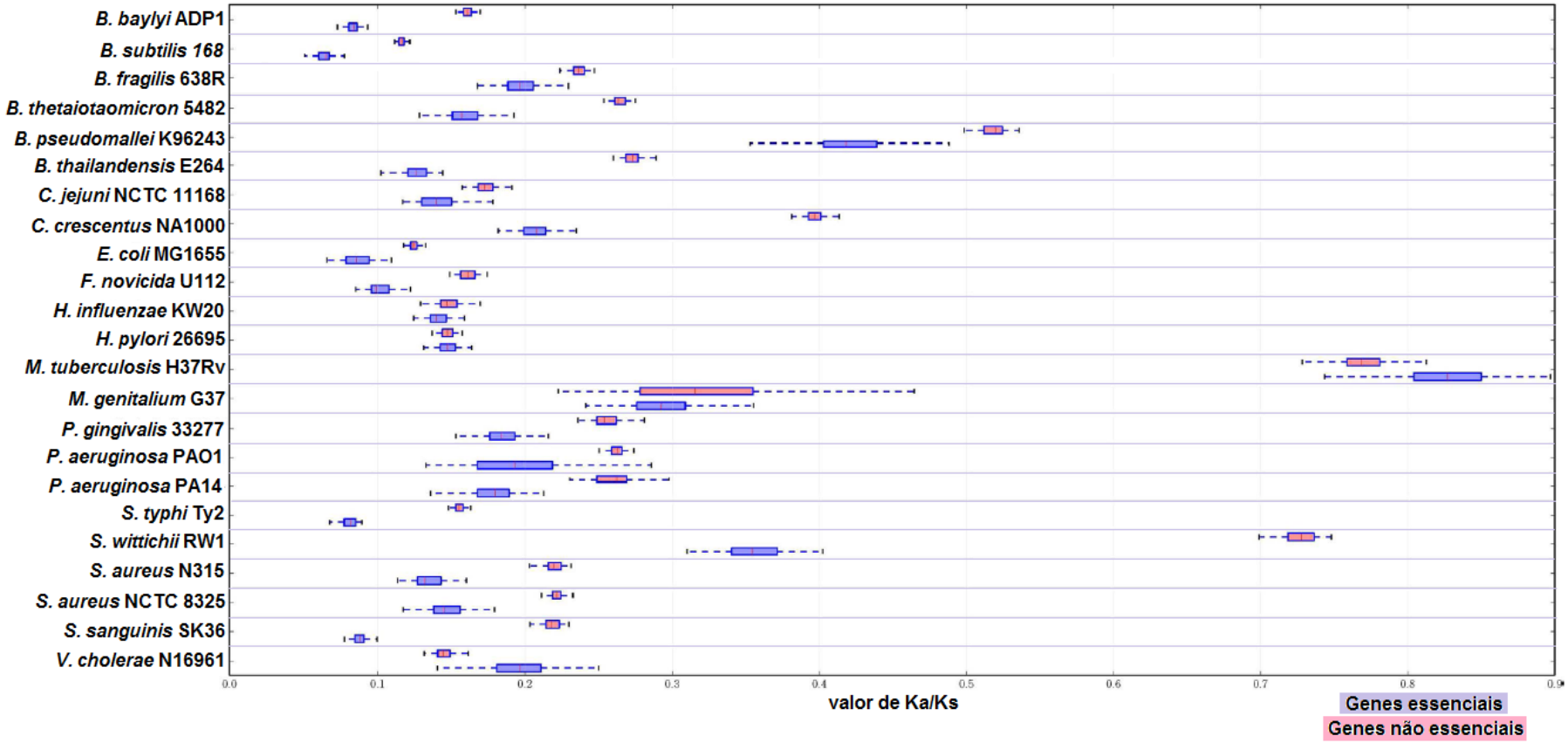


Figura 4 - Par de boxplot comparando a distribuição das médias de Ka/Ks dos genes essenciais em relação à distribuição das médias dos genes não essenciais. Fonte: adaptado a partir da referência [47].

Em termos de conservação ao longo das espécies, um subgrupo dentro do grupo de genes essenciais em um determinado organismo apresentará ortólogos em outros indivíduos mesmo que filogeneticamente distantes. Isso porque, uma parte dos genes essenciais está relacionada à maquinaria celular básica que é amplamente conservada em procariotos e eucariotos. Entretanto, as análises de ortologia entre as listas de genes essenciais provenientes de organismos diferentes têm mostrado as diferenças de requerimentos e as particularidades de cada organismo. Dos 499 genes essenciais em *Acinetobacter baylyi*, 125 são também essenciais em *B. subtilis*, *E. coli* e *P. aeruginosa* e 69 genes são essenciais exclusivos de *A. baylyi*, dos quais 82% não possui função determinada [28]. Dos 271 genes essenciais em *B. subtilis*, 150 são também essenciais em *E. coli*, enquanto 67 são essenciais exclusivos de *B. subtilis* e 86 exclusivos de *E. coli* [27]. O estudo de *Porphyromonas gingivalis* comparou o grupo de genes essenciais obtidos para esta bactéria com os dados de essencialidade a partir de todas as bactérias disponíveis no DEG (*Database of Essential Genes*) em 2011 [35]. Dos 463 genes essenciais de *P. gingivalis*, 364 genes apresentaram ortólogos essenciais em pelo menos uma espécie presente no banco de dados. Além disso, 93% dos genes essenciais pertencem ao *core* genoma de 1.476 genes obtidos a partir da comparação de 10 linhagens de *P. gingivalis* [35]. Devido às características funcionais e a grande similaridade entre genomas de uma mesma espécie, esta sobreposição entre os genes essenciais e o *core* genoma da espécie é esperada. Entretanto, à medida que organismos filogeneticamente mais distantes são incorporados às análises de ortologia, o número de genes ortólogos encontrado é reduzido. Assim, somente 61% dos genes essenciais de *E. coli* são universalmente conservados em Gama-Proteobacteria, enquanto apenas 34% dos genes essenciais de *B. subtilis* são conservados em todos os indivíduos do filo Firmicutes [48]. Dessa forma, alguns autores têm proposto o conceito de persistência na predição de essencialidade (genes compartilhados entre a maioria dos genomas em questão) [48], que será discutido na próxima seção.

3.2. Comparando a organização dos genes essenciais

Os operons compreendem uma das características genômicas conservadas entre os procariotos. Geralmente, mais de 60% dos genes no genoma são co-transcritos a partir de operons [49,50]. O modelo da co-regulação dos genes em operons explica os benefícios da proximidade dos genes e da manutenção dos operons e em contraste com a teoria do

operon egoísta, o modelo da co-regulação assume que os genes cuja regulação é crítica para a função (como os genes essenciais) iriam predominar em operons [51]. De acordo com esta premissa, dois estudos demonstraram o enriquecimento de genes essenciais determinados experimentalmente [52] em estruturas policistrônicas [48]. Este sinal se mantém forte mesmo quando genes codificadores de proteínas ribossomais (que constituem operons de genes essenciais grandes e conservados em procariotos) são removidos da análise [48]. Além da tendência a se apresentarem em operons, foi mostrado que os genes essenciais tendem a estar mais próximos uns dos outros no cromossomo quando comparados aos genes não essenciais, assim como operons contendo genes essenciais também tendem a se agrupar [48,53]. A organização genômica de *E. coli* tem sido relacionada ao tempo de expressão dos genes e dosagem dos genes [54]. O desequilíbrio na concentração dos complexos proteicos pode resultar em redução da adaptabilidade da célula [55]. Em *Helicobacter pylori*, por exemplo, genes codificando proteínas que interagem formando complexos estão localizados mais próximos entre si do que em relação aos genes cujos produtos não formam complexos proteicos [56]. Entretanto, a maioria dos operons não transcreve e traduz proteínas participantes de complexos e o agrupamento dos genes é provavelmente mantido evolutivamente para preservar padrões de expressão similares [57], necessários tanto para genes funcionalmente relacionados (e.g. de uma mesma via metabólica), quanto para genes não relacionados funcionalmente, mas que são co-transcritos. De fato, os genes poderiam estar sob pressões seletivas para localizarem-se em operons ou para serem mantidos próximos entre si a fim de garantir a regulação “afinada” da transcrição, da expressão e da montagem de complexos (quando for o caso, como para as proteínas ribossomais) para múltiplos genes.

Embora um grande número de evidências mostre que a arquitetura dos genomas procarióticos não é aleatória, a organização do genoma é bastante dinâmica ao longo da escala evolutiva. Entretanto, as forças seletivas que moldam os genomas ainda não são completamente entendidas. Em organismos filogeneticamente muito próximos, a ordem dos genes é normalmente preservada [10]. Em contraste, a conservação da ordem dos genes decresce com o aumento da distância filogenética [10]; ainda assim, mesmo entre organismos distantes alguma conservação é encontrada para proteínas que interagem entre si, tais como as proteínas ribossomais [58,59]. Desse modo, poucos operons são compartilhados entre um amplo espectro de organismos [60] e mesmo aqueles operons

ribossomais conservados são encontrados em diferentes arranjos entre os genomas bacterianos e de arqueias [60,61]. Interessantemente, um estudo avaliou a ordenação dos genes essenciais em operons em 13 organismos que possuíam informação de essencialidade e revelou que os genes essenciais predominam na metade proximal (em direção à extremidade 5´) dos operons (Tabela 1) [62]. O mesmo estudo avaliou a ordenação dos genes essenciais em *M. leprae* a partir da extrapolação dos genes essenciais de *M. tuberculosis*. Enquanto os genes essenciais preditos em *M. leprae* predominaram na metade proximal dos operons, os pseudogenes predominaram na metade distal (em direção a extremidade 3´) do operon, sugerindo que os genes funcionalmente mais importantes para a célula iriam se localizar à montante (à frente) de genes menos importantes na organização dos operons em *M. leprae* [62]. Aproveitando-se dessa ideia ainda não amplamente abordada na literatura, uma das análises desta tese discute sobre a ordenação dos genes essenciais nos operons e as implicações dessa ordem nas relações entre função, regulação, expressão e adaptabilidade dos operons.

Tabela 1 - Ordenação dos genes essenciais em operons.

Organismo	Genes	Genes Essenciais	Genes na metade 5'	Genes na metade 3'
Acinetobacter baylyi ADP1	3407	499	97	69
Bacillus subtilis 168	4354	271	46	59
Escherichia coli K12	4320	712	144	134
Francisella novicida U112	1767	392	99	71
Haemophilus influenzae Rd KW20	1738	642	112	120
Helicobacter pylori 26695	1616	323	115	95
Mycobacterium tuberculosis H37Rv	3959	614	219	204
Mycoplasma genitalium G37	518	381	56	51
Mycoplasma pulmonis UAB CTIP	815	310	50	28
Pseudomonas aeruginosa PA14	5964	335	98	94
Salmonella enterica Typhimurium LT2	4541	230	55	56
Staphylococcus aureus N315	2662	302	61	61
Streptococcus pneumoniae R6	2115	244	50	40

Legenda: Um total de 1.202 genes essenciais foram localizados na metade 5´ dos operons enquanto 1.082 genes essenciais ocuparam a extremidade 3´ dos operons, mostrando tendência dos genes essenciais estarem localizados na extremidade proximal (P = 0.006). Tabela adaptada a partir da referência [62].

A organização do cromossomo bacteriano também foi relacionada à indispensabilidade dos genes [63]. Genes essenciais são localizados mais frequentemente na fita líder de replicação quando comparados aos genes não essenciais (Tabela 2) [63,64]. Entretanto, dentre os grupos de genes essenciais determinados experimentalmente em 10 bactérias, somente 10 das 26 categorias funcionais foram significativamente predominantes na fita líder [64]. Essas 10 categorias incluem a maquinaria básica e outras atividades importantes como divisão e síntese da parede celular: “Tradução, biossíntese e estrutura ribossomal” (J), “Transcrição” (K), “Replicação, recombinação e reparo” (L), “Controle do ciclo celular, divisão celular e partição do cromossomo” (D), “Biogênese da membrana celular, envelope e parede celular” (M), “Modificação pos-traducional, conversão de proteínas e chaperonas” (O), “Produção e conversão de energia” (C), “Transporte e metabolismo de carboidratos” (G), “Transporte e metabolismo de aminoácidos” (E) e “Transporte e metabolismo de nucleotídeos” (F) [64]. Por isso, foi sugerido que as funções mais importantes guiariam a organização e a distribuição dos genes entre as fitas do cromossomo em Bacteria [63,64]. Além dos genes essenciais, trabalhos anteriores mostraram que genes altamente expressos em Bacteria são predominantes na fita líder do cromossomo [65]. Uma explicação proposta para esse padrão de localização entre fitas para os genes altamente expressos está relacionada à presença concomitante da DNA polimerase e RNA polimerase no cromossomo. Simultaneamente à replicação do DNA, as bactérias realizam a transcrição e tradução das proteínas. Em determinado momento, a forquilha de replicação poderá encontrar com o aparato de transcrição e assim, as duas enzimas (DNA polimerase e RNA polimerase) irão se encontrar na mesma região cromossômica, podendo atrasar a replicação. Contudo, se o gene estiver presente na fita líder da replicação, significa que as duas enzimas irão prosseguir a síntese de ácidos nucleicos no mesmo sentido (de 5´ para 3´), evitando colisões cabeça-cabeça e consequentemente, evitando o atraso da replicação.

Tabela 2 - Participação dos genes essenciais e não essenciais na fita líder e na fita atrasada do cromossomo para 10 genomas bacterianos.

Organismo	Número de Genes Essenciais			Número de Genes Não Essenciais			Valor de *P*
	Fíta Lider	Fíta Atrasada	Total	Fíta Líder	Fíta Atrasada	Total	
Acinetobacter baylyi	312 (62.5%)	187 (37.5%)	499	1707 (60.4%)	1119 (39.6%)	2826	0.398046
Bacillus subtilis	252 (93.0%)	19 (7.0%)	271	2792 (72.8%)	1042 (27.2%)	3834	0
Escherichia coli	387 (63.4%)	223 (36.6%)	610	1861 (52.8%)	1661 (47.2%)	3522	0.000001
Haemophilus influenzae	366 (57.0%)	276 (43.0%)	642	547 (53.9%)	468 (46.1%)	1015	0.223891
Helicobacter pylori	200 (61.9%)	123 (38.1%)	323	721 (57.5%)	532 (42.5%)	1253	0.163852
Mycoplasma genitalium	317 (83.6%)	62 (16.4%)	379	68 (69.4%)	30 (30.6%)	98	0.002384
Mycoplasma pulmonis	208 (67.1%)	102 (32.9%)	310	276 (58.5%)	196 (41.5%)	472	0.016082
Mycobacterium tuberculosis	449 (73.1%)	165 (26.9%)	614	1895 (56.1%)	1480 (43.9%)	3375	0
Staphylococcus aureus	273 (90.4%)	29 (9.6%)	302	1663 (72.7%)	623 (27.3%)	2286	0
Salmonella typhimurium	165 (71.7%)	65 (28.3%)	230	2429 (57.9%)	1766 (42.1%)	4195	0.000026

Legenda: Tabela adaptada a partir da referência [64].

O mesmo estudo que mostrou o predomínio dos genes altamente expressos na fita líder incluiu os dados de essencialidade e o índice de adaptação de códon dos genes na avaliação seguinte. Essa análise mostrou que tanto os genes essenciais altamente expressos quanto os genes essenciais não altamente expressos predominam significativamente na fita líder [63,66]. Assim, os autores sugeriram que é a essencialidade e não a taxa de expressão que exerce uma influência mais forte para a seleção dos genes na fita líder, evitando a presença de transcritos truncados que poderiam levar a síntese de produtos tóxicos ou deletérios para a célula.

Em contraste, outra teoria mostrou uma correlação forte entre o nível de expressão e a preferência por fita para operons contendo genes altamente expressos, independente da essencialidade dos genes [67]. De acordo com esta teoria, as colisões seriam deletérias porque as interrupções na expressão de qualquer gene (seja ele essencial ou não essencial) levariam ao desequilíbrio do crescimento celular. Neste caso, a organização do cromossomo seria relacionada à importância do produto gênico para a adaptabilidade da célula, independente do status de essencialidade. Seguindo este raciocínio, alguns estudos propuseram o conceito de persistência do gene para o qual, aqueles genes conservados na maioria dos genomas e dispersos ao longo da árvore filogenética seriam os verdadeiros genes essenciais [33,48,68].

As análises baseadas em persistência permitem a identificação de genes essenciais que são frequentemente perdidos na determinação experimental de genes essenciais, como por exemplo, genes de reparo de DNA [33,48]. Assim, o grupo de genes persistentes é enriquecido em genes envolvidos na manutenção celular e resposta ao estresse [68]. Estes genes, apesar de dispensáveis em condições ótimas de laboratório, seriam essenciais para a sobrevivência em condições naturais [68]. Além disso, genes essenciais e genes persistentes não essenciais compartilham características comuns, tais como alta conservação de sequência e taxas de expressão (preditas pelo índice de adaptação de códon), localização preferencial na fita líder do cromossomo (minimizando o risco de colisões entre a DNA polimerase e a RNA polimerase) [69], e a tendência para estar em operons [48]. Por isso a persistência tem sido relacionada com a organização do genoma [68]. Deste modo, todos os estudos mencionados acima demonstram que as características de função, conservação e organização dos genes estão submetidas a pressões seletivas que mantêm a estrutura dos genomas. O conhecimento dessas características será útil para o desenho, experimentação

e manufatura de células artificiais e nos fornecerá as bases científicas necessárias para o campo, ainda em desbravamento, da biologia sintética.

Embora descobertas importantes tenham sido relatadas por meio de estudos *in silico* (genômica comparativa) e estudos experimentais de essencialidade dos genes em *E. coli*, muitos progressos foram alcançados por meio dos métodos de sequenciamento de segunda geração e de mutações em larga escala que têm permitido a avaliação de espécies distantemente relacionadas com alta resolução. Nesta tese, foram analisados os dados de essencialidade de 17 estudos experimentais *in vitro* conduzidos para 16 organismos. Também foram incorporados dados de essencialidade *in vivo* para os organismos selecionados com experimentos *in vitro*. Para dois organismos, três estudos experimentais *in vivo* foram encontrados. Em contraste com estudos anteriores, este estudo incluiu os dados de essencialidade da única arqueia com determinação experimental até o momento [70]. Portanto, nossas análises compreendem os dois domínios procarióticos da vida.

OBJETIVO GERAL

Analisar comparativamente, por meio de métodos computacionais, os grupos de genes essenciais experimentalmente determinados em procariotos de modo a compreender as características únicas ou compartilhadas de função, conservação e organização dos genes essenciais entre os organismos selecionados.

OBJETIVOS ESPECÍFICOS

- Identificar o conjunto de genes essenciais exclusivos de cada organismo selecionado;
- Identificar o conjunto de genes essenciais compartilhados entre todos os organismos selecionados;
- Identificar o grupo de funções com a maior participação dos genes essenciais nos procariotos;
- Avaliar a contribuição dos genes essenciais na organização do genoma, ou seja, a participação na estrutura dos operons ou como genes monocistrônicos.

MÉTODOS

1. Obtenção dos dados

1.1. Genes essenciais

As listas de genes essenciais foram selecionadas após cautelosa avaliação de todos os artigos publicados apresentando dados de essencialidade. Durante a avaliação destes artigos, os seguintes critérios de inclusão foram considerados para a seleção dos trabalhos:

a) Identificação experimental *in vitro* dos genes essenciais;
b) Determinação dos genes essenciais de forma sistemática, ou seja, estudo abrangendo todo o cromossomo procariótico. Quando estratégias de deleção gene a gene foram conduzidas, pelo menos 80% dos genes codificadores de proteínas deveriam ter sido alvos de deleção;
c) Quando estratégias de mutação por transposons foram empregadas, os estudos deveriam ter alcançado o ponto de saturação ou indicarem a proximidade do estado de saturação das transposições;
d) Interpretação do estudo baseada em letalidade;
e) Dados de essencialidade foram considerados para genes codificadores de proteínas.

Seguindo estes critérios, 17 estudos foram selecionados, compreendendo 16 organismos, sendo 15 bactérias e uma arquéia (Tabela 3). Dentre os estudos considerados, dois apresentaram exceções em relação aos critérios indicados. A interpretação de essencialidade no estudo de *S. enterica* sorovar typhimurium SL1344 [37] foi baseada na redução do valor adaptativo celular ao invés de letalidade. Este estudo foi incluído em nossa análise porque 95% dos genes requeridos corroboram os dados de genes essenciais baseados em letalidade de um estudo anterior do mesmo grupo [71]. No estudo de *B. subtilis* 168 [26], os genes essenciais foram determinados pela combinação de dados experimentais com dados preditos. Este estudo foi considerado porque somente 4% (184/4100) dos genes codificadores de proteínas presentes no genoma foram preditos.

1.2. Genomas

Os dados relacionados às informações do genoma, tais como, identificadores dos genes, coordenadas gênicas, nome dos genes, informação de fita e as sequências dos genes e das proteínas, foram extraídos a partir dos arquivos Genbank [72] de cada organismo.

1.3. Predição dos operons

A predição da estrutura dos operons para cada organismo foi obtida a partir do banco de dados DOOR2 [73]. Este banco de dados tem sido indicado como o repositório mais confiável para a predição de operons procarióticos [74].

1.4. Arquitetura de domínios proteicos

As arquiteturas dos domínios das sequências de proteínas foram analisadas por meio do programa HMMER versão 3.0 [75], utilizando o parâmetro de e-value ≤ 0.01, e o banco de dados PFAM versão 27.0 [76].

1.5. Análises

Os dados foram processados por meio de *scripts* escritos em Perl, R (www.r-project.org) e *shell*. Para cada organismo, três listas foram obtidas: a lista de genes codificadores de proteínas (a partir do arquivo genbank), a lista de genes essenciais (a partir do trabalho experimental) e a lista de genes policistrônicos (a partir do DOOR2). Estas três listas foram analisadas de modo a extrair a informação sobre quais genes eram monocistrônicos (essenciais e não essenciais) e policistrônicos (essenciais e não essenciais), bem como a informação da fração de genes essenciais no genoma de cada organismo.

2. Análises de homologia

2.1. Obtenção dos COGs/NOGs

Os dados de homologia foram obtidos a partir do banco de dados EGGNOG versão 4.0 [77]. As sequências de proteínas foram mapeadas ao EGGNOG (2031 espécies core-

periféricas) por meio da ferramenta BLAST [78]. Os NOGs (*nonsupervised orthogous groups*) são uma extensão dos COGs (*cluster of orthologous groups*) manualmente anotados [77,79]. Uma vez que as espécies selecionadas neste trabalho ou os organismos proximamente relacionados estão presentes no EGGNOG, parâmetros rigorosos foram utilizados para a identificação dos homológos por meio do BLAST: e-value ≤ 10^{-10} e $S \geq 60\%$, para o qual S representa a cobertura da sequência mais curta (*query* ou *subject*).

2.2. Conservação dos genes e análise de diversidade dos grupos essenciais

Um índice de persistência para cada NOG foi calculado como previamente descrito [68] como a fração das espécies com pelo menos um homólogo naquele NOG. Os padrões de presença ou ausência de genes essenciais a partir de uma dada espécie em um NOG em cada condição foram usados para criar uma matriz Booleana que foi usada para realizar uma análise de MCA (*Multiple Correspondence Analysis*) por meio do pacote FACTORMINE [80].

2.3. Enriquecimento funcional, famílias gênicas e genes universalmente conservados

A análise de enriquecimento da categoria funcional foi calculada por meio do teste exato de Fisher ($P < 0.05$). Genes anotados para o mesmo NOG em uma espécie foram considerados como parte da mesma família gênica e, portanto, presentes em uma família multigênica (dois ou mais genes). Genes cujas anotações de NOG foram exclusivas – não compartilhadas com outros genes na espécie – foram considerados como os únicos representantes da família gênica e, portanto, presentes em uma família monogênica.

2.4. Transferência horizontal dos genes

Os genes envolvidos em eventos de transferência gênica lateral foram identificados por meio do banco de dados DARKHORSE [81]. Todas as espécies, exceto duas (*S. enterica* typhimurium SL1344 e *B. fragilis* 638R), apresentam informação sobre transferência horizontal no banco de dados. O algoritmo usado pelo DARKHORSE baseia-se na análise estatística dos genomas de bactéria e arqueias para a identificação de proteínas filogeneticamente “atípicas” [81].

3. Organização do genoma

3.1. Presença de genes essenciais em operons

Para cada organismo selecionado, a lista de genes essenciais foi obtida a partir do respectivo trabalho experimental e a lista de todos os genes codificadores de proteínas, a partir do Genbank. Com essas informações, foi obtida uma tabela de contingência 2x2 com as seguintes categorias:

a) genes policistrônicos essenciais;

b) genes policistrônicos não essenciais;

c) genes monocistrônicos essenciais;

d) genes monocistrônicos não essenciais.

As associações entre as quatro categorias foram avaliadas estatisticamente por meio do teste exato de Fisher ($P \leq 0.05$). A mesma análise de associação foi realizada por meio de simulações. Dez mil arquivos foram simulados para cada um dos organismos estudados, preservando-se o número de genes essenciais, o número de operons e o número de genes em cada operon, porém, aleatorizando quais genes estavam presentes nos operons. A associação entre as categorias foi então analisada para cada um dos 10.000 arquivos simulados. As análises estatísticas foram realizadas no ambiente estatístico R.

3.2. Posição dos genes essenciais em operons

Em cada um dos organismos selecionados foi avaliada a frequência do tamanho dos operons no genoma. O tamanho do operon foi definido pelo número de genes na sua estrutura. Por exemplo, operons de tamanho dois apresentam dois genes na sua estrutura enquanto operons de tamanho três apresentam três genes, e assim por diante. Para isso, o número de operons com cada um dos tamanhos possíveis para cada genoma foi contabilizado e a frequência em que se faziam presentes para cada organismo determinada. Para todos os organismos, operons de tamanho dois e três foram os mais frequentes no genoma. Portanto, a análise seguinte foi conduzida apenas para os operons mais representativos nos genomas. Foi avaliado se havia posição preferencial dos genes essenciais na estrutura dos operons. Para isso, apenas os operons contendo pelo menos um gene essencial em sua estrutura foram considerados. A posição de todos os genes essenciais em cada operon de tamanho dois e de tamanho três foram identificadas. Com a

informação do número de genes essenciais ocupando as primeira e a segunda posições em operons de tamanho dois, por exemplo, foi possível avaliar as divergências entre a frequência esperada e observada para estes operons por meio do teste de Chi-quadrado ($P \leq 0.01$).

RESULTADOS E DISCUSSÃO

1. O número de genes essenciais não apresenta correlação com o tamanho dos genomas procarióticos

A seleção dos estudos de determinação experimental de genes essenciais resultou em uma lista de 17 experimentos *in vitro* para 16 organismos, compreendendo seis diferentes filos (Tabela 3). A análise quantitativa e qualitativa das 17 listas de genes essenciais revelou diferenças relevantes de ordem técnica e biológica. Em relação às diferenças técnicas, os estudos baseados em deleções gene a gene para a identificação dos genes essenciais conduzidos para *S. sanguinis*, *E. coli* e *B. subtilis* apresentaram conjuntos de genes essenciais menores em relação aos demais trabalhos selecionados (Tabela 3 e Figura 5). Esta técnica é considerada como o padrão ouro para a identificação de genes essenciais por realizar a deleção completa de cada gene e ser menos passível a erros do que as técnicas de mutagênese. Em relação às diferenças quantitativas por motivo biológico, por exemplo, os maiores conjuntos de genes essenciais foram encontrados para as espécies *Methanococcus maripaludis* e *M. tuberculosis*. Em contraste com a maioria dos organismos selecionados, as duas espécies mencionadas dependem de um ambiente anaeróbico para o crescimento, apresentam requerimentos nutricionais complexos e meios de cultivo nutricionalmente ideais ainda não estão disponíveis para estas espécies. Desse modo, o número de genes requeridos para o crescimento e manutenção do organismo *in vitro* em um meio de cultivo subótimo é maximizado.

A análise comparativa do número de genes essenciais com o número de genes codificadores de proteínas para cada organismo mostrou a ausência de correlação entre o tamanho do conjunto essencial e o tamanho do genoma (Figura 5). Enquanto o número de genes codificadores de proteína varia de um mínimo de 475 para *M. genitalium* até o máximo de 5.727 para *Burkholderia pseudomallei*, o número de genes essenciais não aumenta proporcionalmente nestas espécies. Assim, o menor conjunto de genes essenciais é encontrado para *S. sanguinis* (218 genes) e o maior, para *M. tuberculosis* (774 genes). Sob condições estáveis em laboratório (crescimento em meio rico e sem estresse) e avaliando a inativação de um gene por vez, a função de um gene removido ou mutado pode ser

compensada por um gene funcionalmente redundante (homólogo ou não). Nesta situação, para a maioria dos estudos, ambos os genes serão considerados não essenciais baseado na interpretação atual de essencialidade. Devido à redundância de genes, em organismos adaptados a um grande número de ambientes e com genomas grandes – como *E. coli* e *B. subtilis*, ambos apresentando 34,5% de genes parálogos – menos de 10% dos seus genes são considerados essenciais. Por outro lado, *M. genitalium* (próximo a um genoma mínimo e com apenas 6% de genes parálogos), apresenta 80% do seu genoma essencial mesmo quando cultivado em meio rico (Figura 5, Anexo 2).

Aparentemente, os conjuntos de genes essenciais são mais limitados nos procariotos quando comparados aos eucariotos, nos quais a fração de genes essenciais parece ser mais proporcional ao tamanho do genoma. Ensaios de identificação de genes essenciais em dois eucariotos, *Saccharomyces cerevisiae* e *Schizosaccharomyces pombe*, apresentou de 1.000 a 1.300 genes essenciais para o crescimento destes organismos [82,83]. Enquanto os tamanhos dos genomas destes fungos (>5.000 genes) são similares aos maiores genomas bacterianos (e.g. *B. pseudomallei*), os conjuntos de genes essenciais são aproximadamente 1,4 vezes maiores que os maiores conjuntos de genes essenciais em procariotos (isto é, *M. tuberculosis* com 774 genes essenciais). Avaliações tão profundas e sistemáticas para determinação de genes essenciais ainda não são disponíveis para eucariotos multicelulares. Assim, à medida que um maior número de estudos sistemáticos de identificação de genes essenciais esteja disponível, tanto para procariotos como para eucariotos, possivelmente poderemos entender mais claramente as tendências na proporção dos conjuntos dos genes essenciais em relação ao tamanho dos genomas, assim como suas implicações, nestes domínios da vida.

Tabela 3 - Trabalhos experimentais selecionados para o presente estudo de tese.

Organismo	Genes Essenciais	CDS[1]	% Genes Essenciais	Meio de Cultura	Estratégia (nome original como descrito no artigo)	Referência
Mycoplasma genitalium G37	382	475	80.40%	Rico	Global transposon mutagenesis + Sanger sequencing	[22]
Mycoplasma pulmonis CT	310	782	39.60%	Rico	Global transposon mutagenesis + Sanger sequencing	[84]
Francisella novicida tularensis U112	396	1719	23.00%	Rico	Sequence-defined transposon mutant library + Sanger sequencing	[85]
Methanococcus maripaludis S2	526	1722	30.50%	Rico	Saturation mutagenesis technique + Illumina sequencing (TnSeq)	[70]
Methanococcus maripaludis S2	664	1722	38.50%	Mínimo	Saturation mutagenesis technique + Illumina sequencing (TnSeq)	[70]
Porphyromonas gingivalis ATCC 33277	463	2090	22.10%	Rico	Global transposon mutagenesis + Illumina sequencing (TnSeq)	[35]
Streptococcus sanguinis SK36	218	2270	9.60%	Rico	Systematic gene replacement	[29]
Acinetobacter baylyi ADP1	499	3307	15.10%	Mínimo	Single-gene-deletion	[28]
Caulobacter crescentus NA1000	480	3877	12.40%	Rico	Hyper-saturated transposon mutagenesis + Illumina sequencing	[86]
Mycobacterium tuberculosis H37Rv	774	4018	19.20%	Mínimo	High-density transposon mutagenesis + Illumina sequencing	[87]
Escherichia coli K12	303	4145	7.30%	Rico	In-frame single gene deletions	[27]
Bacillus subtilis 168	271	4176	6.40%	Rico	Gene-by-gene inactivation	[26]
Bacteroides fragilis 638R	550	4290	12.80%	Rico	Transposon delivery vetor + Illumina sequencing	[88]
Salmonella enterica typhi Ty2	356	4370	8.10%	Rico	Transposon directed insertion sequencing site (TRADIS)	[71]
Salmonella enterica typhimurium SL1344	353	4446	7.90%	Rico	Transposon directed insertion sequencing site (TRADIS)	[37]
Burkholderia thailandensis E264	406	5632	7.20%	Rico	Saturation level transposon mutagenesis + Illumina sequencing (TnSeq)	[89]
Burkholderia pseudomallei K96243	505	5727	8.80%	Rico	Transposon directed insertion sequencing site (TRADIS)	[90]

Legenda: [1] CDS, Coding sequences ou genes codificadores de proteínas. Fonte: tabela adaptada a partir da referência [4].

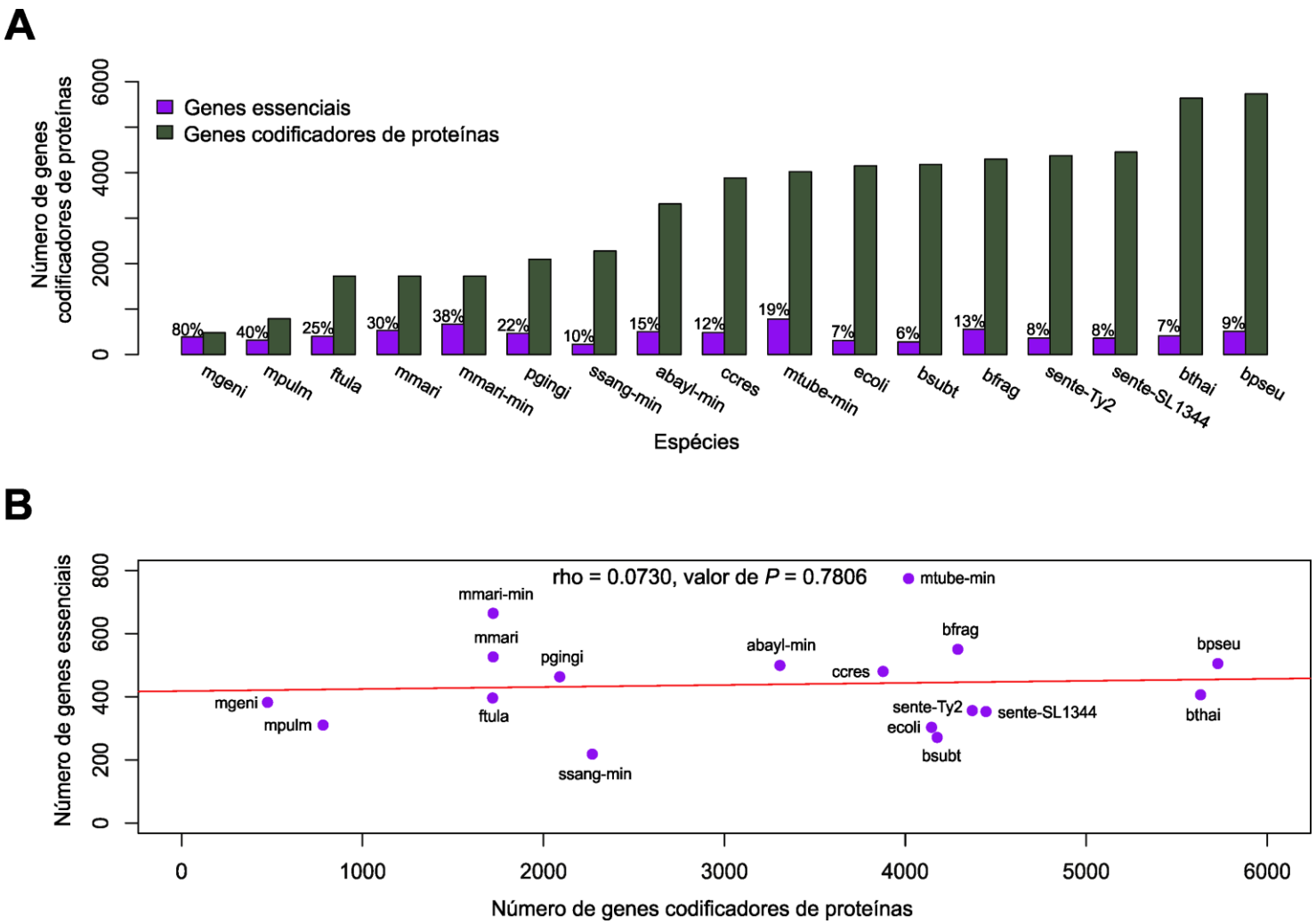


Figura 5 - Genes essenciais obtidos a partir de 17 experimentos e sua correlação com o número de genes codificadores no genoma. A) Porcentagem dos genes essenciais em cada genoma, B) Correlação entre o número de genes essenciais e o número de genes codificadores de proteínas.
Legenda: abayl-min (Acinetobacter baylyi ADP1), bfrag (Bacteroides fragilis 638R), bpseu (Burkholderia pseudomallei K96243), bsubt (Bacillus subtilis 168), bthai (Burkholderia thailandensis E264), ccres (Caulobacter crescentus NA1000), ecoli (Escherichia coli K-12), ftula (Francisella tularensis novicida U112), mgeni (Mycoplasma genitalium G37), mmari (Methanococcus maripaludis S2), mmari-min (M. maripaludis S2, meio mínimo), mpulm (Mycoplasma pulmonis UAB CTIP), mtube-min (Mycobacterium tuberculosis H37Rv), pging (Porphyromonas gingivalis ATCC 33277), sente-SL1344 (Salmonella entérica typhimurium SL1344), sente-Ty2 (Salmonella entérica typhi Ty2), ssang (Streptococcus sanguinis SK36).

Fonte: ilustração adaptada a partir da referência [4].

2. Categorias funcionais relacionadas à proliferação celular são enriquecidas em genes essenciais

A vasta maioria dos estudos experimentais de essencialidade tem sido conduzida para o crescimento dos organismos em meio rico. O crescimento em um meio considerado ideal permite que o organismo possa crescer sem a necessidade de usar várias vias biosintéticas. Assim, em condição de crescimento ideal, os estudos têm apresentado os conjuntos de genes essenciais enriquecidos para os genes que constituem a maquinaria básica para a sobrevivência, tais como a replicação do DNA e a síntese de proteínas. Esta observação foi confirmada na análise de enriquecimento de genes essenciais para as categorias funcionais encontradas em cada genoma. Houve o enriquecimento de genes essenciais na categoria funcional de “Tradução, biogênese e estrutura ribossomal” (J) em todos os experimentos *in vitro* de meio rico e mínimo (Figura 6). Como esperado, independente do meio no qual se encontre, nenhum organismo é capaz de sobreviver e multiplicar-se sem a capacidade de produzir as proteínas da própria estrutura de tradução e/ou realizar a montagem do complexo de tradução. Outras categorias funcionais que apresentaram enriquecimento de genes essenciais na maioria das espécies foram relacionadas à parede celular, ao metabolismo de lipídeos, ao controle do ciclo e divisão celular e ao metabolismo de coenzimas. Estas funções são discutidas em mais detalhes em seguida.

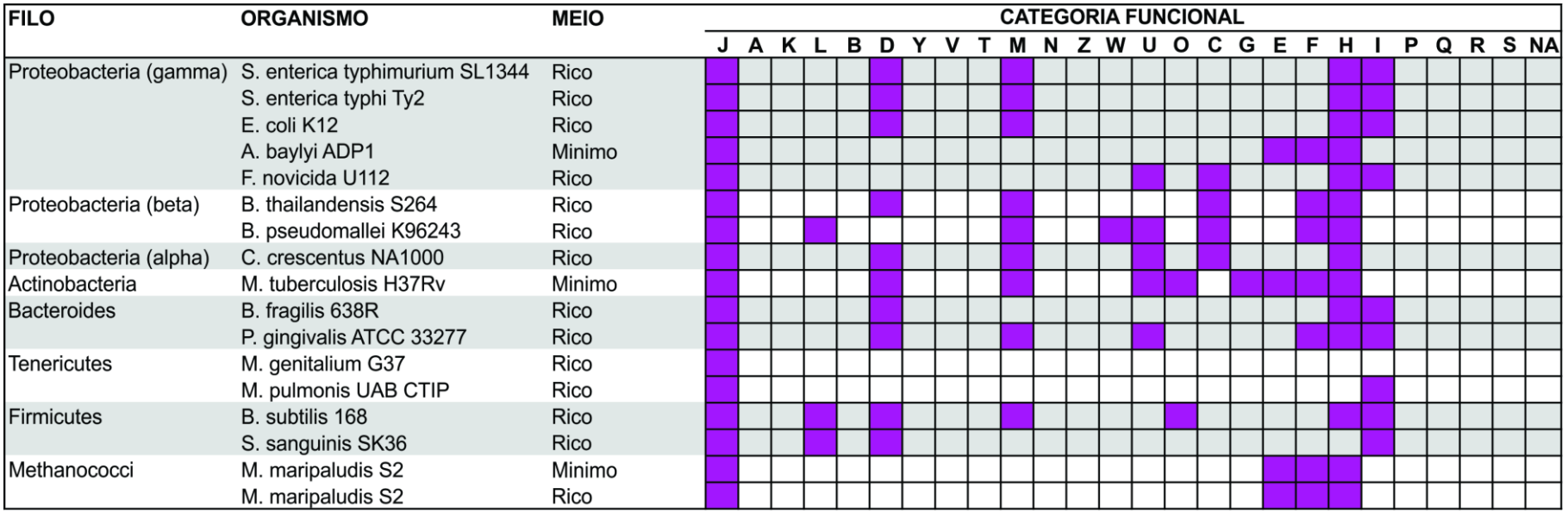


Figura 6 - Categorias funcionais enriquecidas em genes essenciais.
Legenda: quadrados em magenta representam as categorias funcionais enriquecidas no respectivo grupo de genes essenciais (Teste exato de Fisher, $P < 0.05$). Categorias: J – Tradução e biosíntese e estrutura ribossomal; A – Modificação e processamento de RNA; K – Trancrição; L – Replicação, recombinação e reparo; B – Dinâmica e estrutura da cromatina; D – Controle do ciclo celular, divisão celular e partição do cromossomo; Y – Estrutura nuclear; V – Mecanismos de defesa; T – mecanismos de transdução de sinal; M – Biosíntese da parede celular, membrana e do envelope; N – Motilidade celular; Z – Citoesqueleto; W – Estrutura extracelular; U – Tráfego intracelular, secreção e transporte vesicular; O – Modificação pós-traducional, turnover de proteínas e chaperonas; C – Produção e conversão de energia; G – Transporte e metabolismo de carboidratos; E – Transporte e metabolismo de aminoácidos; F – Transporte e metabolismo de nucleotídeos; H – Transporte e metabolismo de coenzimas; I – Transporte e metabolismo de lipídeos; P – Transporte e metabolismo de íon inorgânico; Q – Biosíntese, transporte e catabolismo de metabólitos secundários; R – Somente predição da função geral; S – Função desconhecida e NA – não assinado para um COG (cluster of orthologous groups). Fonte: ilustração adaptada a partir da referência [4].

2.1. Parede celular

É importante ressaltar que a análise de enriquecimento dos genes essenciais nas categorias funcionais baseou-se na comparação intraespecífica dos grupos funcionais e não na análise comparativa de enriquecimento dos grupos funcionais entre espécies. Desse modo, baseado na determinação experimental – que é independente de conceitos de ortologia como visto nos trabalhos de predição computacional – funções que estão presentes em poucos organismos e funções desempenhadas em diferentes organismos realizadas por genes não ortológos podem ser detectadas como funcionalmente enriquecidas para genes essenciais nas diversas espécies. Um exemplo considerável está relacionado à parede celular nas diferentes bactérias. A composição da parede celular é bastante distinta entre bactérias gram-positivas e gram-negativas [91]. Além disso, bactérias como os micoplasmas, sequer apresentam parede celular. Por isso, em análises de predição computacional de essencialidade por meio de genômica comparativa entre as espécies – baseadas no conceito de ortologia – os genes relacionados à parede celular não são identificados como essenciais ou enriquecidos para a categoria de síntese da parede celular [33]. Entretanto, a avaliação apresentada nesta tese baseou-se em dado experimental e foi independente da análise de ortólogos entre as diferentes espécies, e assim foi identificado o enriquecimento de genes essenciais para várias espécies na categoria “Biogênese da parede celular, membrana e envelope” (M) independentemente do tipo de parede celular.

2.2. Metabolismo de lipídeos

A categoria de “Metabolismo e transporte de lipídeos” (I) apresenta diferenças significativas entre os domínios Bacteria e Archaea. Os isoprenóides ou terpenóides são elementos importantes da membrana e da parede celular procariótica [92]. Fosfolipídeos também são críticos na composição da membrana plasmática e sua biossíntese é amplamente conservada em bactérias [93]. Analisando a essencialidade dos genes individualmente para cada espécie, nós encontramos que as enzimas fosfatidato citidil-transferase (EC 2.7.7.41) e CDP-diacilglicerol-glicerol-3-fosfato 3-fosfatidiltransferase (EC 2.7.8.5), ambas envolvidas no metabolismo de glicerofosfolipídeos (KEGG ec00564) foram conservadas em quase todas as espécies (Tabela Suplementar 1, material online). Por outro lado, o papel dos ácidos graxos em Archaea permanece controverso, uma vez que a membrana em arqueias depende da síntese de isoprenóides pela via do mevalonato (KEGG

M00095) [94]. A via do mevalonato compreende os primeiros passos (a partir da acetil-CoA até o isopentenil-pirofosfato) na biosíntese de terpenóides (KEGG ec00900) em arquéias, fungos e metazoários, enquanto bactérias e protozoários (filo Apicomplexa) realizam estes passos até o isopentenil-pirofosfato por meio da via do metileritritol 4-fosfato (KEGG M00096) [95].

Na arqueia *M. maripaludis*, as enzimas da via do mevalonato (MMP1212, MMP1211, MMP0087, MMP1335) são essenciais, enquanto aquelas a partir da via do metileritritol 4-fosfato são essenciais nas bactérias. Surpreendentemente, genes da via do mevalonato (SSA_0338, SSA_0337, SSA_0333, SSA_0335, SSA_0334) são essenciais na bactéria *S. sanguinis*. A essencialidade de genes nessa via foi anteriormente mostrada por um estudo em uma outra espécie do mesmo gênero, *S. pneumoniae* [96]. Em contraste com outras bactérias, *S. sanguinis* utiliza a via do mevalonato para a síntese de lipídeos; esta via foi possivelmente transferida horizontalmente a partir de arqueia ou de células eucarióticas [92,96] e provavelmente substituiu a via do metileritritol 4-fosfato ao longo do tempo. Por fim, a enzima undecaprenil pirofosfato sintase (EC 2.5.1.31) que atua como um transportador de lipídeos para a síntese de peptideoglicanas em bactéria – provável transportador glicosil em Archaea – e está envolvida no passo final da biossíntese de terpenóides (ec00900), é essencial em Archaea e na maioria das bactérias, exceto nos micoplasmas, os quais não apresentam parede celular.

2.3. Controle do ciclo e da divisão celular

O processo de divisão celular é diretamente relacionado ao remodelamento da membrana celular, à divisão do citoplasma e do DNA entre as células filhas e ao aumento da massa celular. Anormalidades nos genes relacionados à divisão celular podem resultar em células filhas com tamanhos assimétricos [97] ou com a divisão celular comprometida [98]. Muitos estudos de divisão celular em bactérias são focados na proteína FtsZ [98], uma proteína conservada na maioria das bactérias e em Euryarchaeota [99], um filo que inclui a arqueia *M. maripaludis* e apresenta um mecanismo de divisão semelhante às bactérias [100]. A proteína FtsZ é importante para a formação do septo [101] e é essencial em todas as espécies, exceto *P. gingivalis* e *M. maripaludis* (Tabela Suplementar 1, material online). Os motivos que levaram a dispensabilidade do gene essencial nestas duas espécies decorrem da interpretação da essencialidade em um estudo e a redundância biológica no outro. Em *P.*

gingivalis, o gene ftsz apresentou apenas duas inserções por transposons em ambas as réplicas técnicas e por isso não foi considerado essencial. *M. maripaludis* apresenta dois genes ftsz (MMP1436 e MMP1500) com arquiteturas de domínios idênticas. Possivelmente, estes genes realizam a mesma função e, portanto, com a deleção ou ruptura de um deles, o outro poderia compensar a função biológica, constituindo um *backup* genético. Dessa forma, o gene foi considerado indispensável no estudo.

2.4. Metabolismo de coenzimas e genes sem função conhecida

A categoria de “Metabolismo e transporte de coenzimas” (H) compreende as vias cujos genes são críticos para outras vias. Os genes requeridos para a síntese de coenzima A (CoA) e biotina (coenzima R/vitamina H), enzimas importantes na oxidação dos ácidos graxos e outras vias metabólicas, são requeridas na maioria das espécies (Tabela Suplementar 1, material online). Além disso, genes envolvidos na produção de nicotinamida (vitamina B3), riboflavina (vitamina B2), folato (vitamina B9) e S-adenosilmetionina são também essenciais. Entretanto, genes que são pobremente caracterizados ou não apresentam anotação de COG representam 28-54% dos conjuntos de genes essenciais. Espera-se que com a melhoria da anotação dos genomas seja possível que outras categorias funcionais enriquecidas em genes essenciais sejam reveladas.

3. Composição dos conjuntos de genes essenciais

3.1. Conservação

Para investigar a conservação dos genes essenciais ao longo da árvore filogenética, os genes codificadores de proteínas dos organismos selecionados para este estudo foram mapeados no banco de dados EGGNOG [77]. Os genes com a mesma anotação de NOG foram considerados homólogos. A conservação dos NOGs nos 2.031 genomas disponíveis no EGGNOG foi avaliada por meio do índice de persistência [68]. Os índices de persistência dos genes essenciais foram comparados aos índices de persistência dos genes não essenciais no mesmo organismo. Essa análise revelou a alta conservação dos genes essenciais em relação aos genes não essenciais em todas as espécies para os experimentos *in vitro* (Figura 7). Estes resultados sugerem a existência de um *core* de genes responsáveis

pelo crescimento em ambos, meio rico e mínimo, apesar da ampla diversidade evolutiva de espécies e dos fenômenos que possam afetar a retenção ou perda de genes, como exemplo, a existência de genes não ortólogos realizando a mesma função [102]. Nossos resultados confirmam as observações realizadas para *E. coli* [48] e *B. subtilis* [68] e estendem as conclusões para todos os grandes grupos de bactérias e uma Archaea. Estes resultados mostram que a alta conservação dos genes essenciais é uma característica comum aos dois domínios procarióticos, Bacteria e Archaea.

A mesma análise de conservação foi realizada para três experimentos *in vivo*, sendo dois de *M. tuberculosis* e um de *F. novicida*, espécies que também fazem parte da seleção de experimentos *in vitro* compreendida nesta tese. Em contraste com o resultado observado para os genes essenciais *in vitro*, os conjuntos de genes requeridos apresentaram índices de persistência mais baixos nos experimentos *in vivo* (Figura 7). Três situações poderiam explicar o resultado para os experimentos *in vivo*. Primeiramente, experimentos *in vivo* geralmente não alcançam a saturação de inserções e por isso, genes altamente conservados que poderiam ser requeridos, não chegam a ser alvos de inserção e não são incluídos na lista de genes requeridos. Segundo, os mutantes obtidos após a transposição são selecionados em um meio de cultura rico antes da inoculação *in vivo* e, portanto, mutantes para genes essenciais *in vitro* são perdidos durante o processo e não chegam a ser testados no experimento *in vivo*, assim não são presentes na lista de genes requeridos. Em terceiro lugar, é provável que em um ambiente bastante distinto de um meio rico ideal, o organismo utilize diferentes estratégias de sobrevivência, contando, por exemplo, com genes de alta variabilidade antigênica empregados durante a evasão do sistema imune hospedeiro. Justamente por esse motivo, embora os genes requeridos *in vivo* apresentem menor taxa de conservação filogenética, eles são considerados de importância médica para o desenvolvimento de novas drogas.

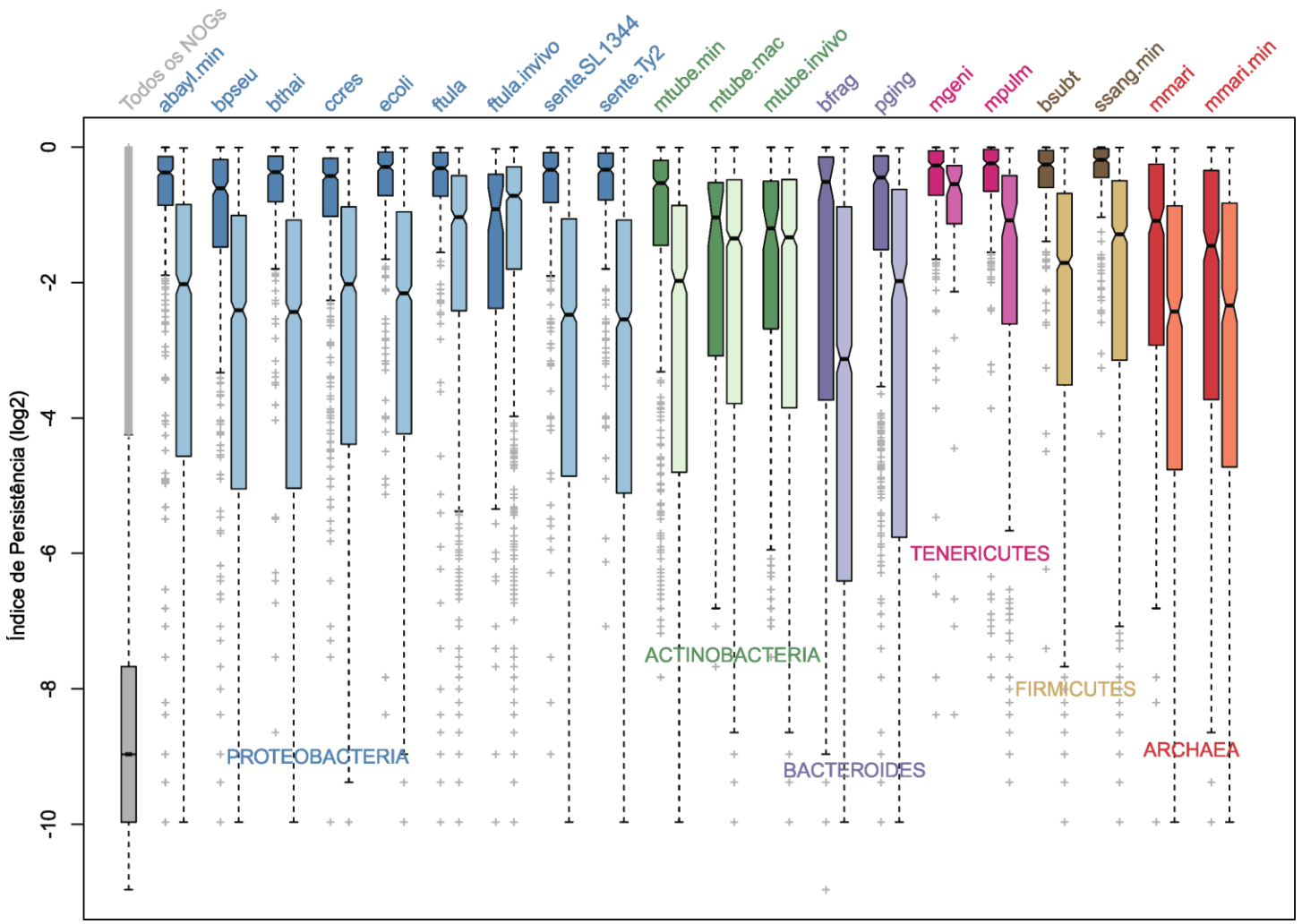


Figura 7 - Conservação dos genes essenciais e não essenciais ao longo de milhares de espécies disponíveis no banco de dados EGGNOG.
Legenda: Para abreviaturas vide Figura 5. Para os experimentos in vivo ftula-invivo (Francisella tularensis novicida U112 em camundongo), mtube-invivo (Mycobacterium tuberculosis H37Rv em camundongo) e mtube-mac (M. tuberculosis H37Rv em macrófago). Para cada organismo o boxplot para a conservação de genes essenciais e não essenciais é apresentado lado a lado. Cores de tons escuros representam genes essenciais e tons claros, os genes não essenciais. Os filos são distinguidos por diferentes cores, sendo Proteobacteria (azul), Actinobacteria (verde), Bacteroides (lilás), Tenericutes (magenta), Firmicutes (marron) e Archaea (vermelho).

Fonte: ilustração adaptada a partir da referência [4].

3.2. Diversidade

Como os repertórios de genes essenciais são qualitativamente distintos entre os organismos e experimentos, uma análise multivariada [80] foi realizada para identificar padrões de variação sistemática nos dados. O dado categórico foi definido como a presença ou ausência de um gene essencial em determinado NOG em cada lista de genes essenciais. A análise multivariada agrupou os organismos pela relação filogenética, assim como, pelo tipo de experimento, permitindo capturar a influência do ambiente (Figura 8). Os conjuntos de genes requeridos *in vivo* em *M. tuberculosis* e *F. novicida* foram agrupados juntos e distanciados dos experimentos *in vitro*. Uma explicação poderia ser relacionada ao repertório de genes requeridos envolvidos com o sistema imune nestes experimentos. Porém, considerando que as estratégias empregadas por estes organismos no hospedeiro são bastante distintas entre si, a aproximação destes dados foi possivelmente relacionada à maior diferença com os grupos de genes essenciais *in vitro*.

Levando em consideração a provável diferença dos conjuntos de genes *in vivo*, o repertório *in vivo* de *M. tuberculosis* foi estudado mais detalhadamente. Primeiramente, os NOGs de genes requeridos exclusivos de *M. tuberculosis* foram identificados com o objetivo de encontrar genes relacionados à infecção e patogenicidade (Tabela Suplementar 2, material online). A lista de NOGs exclusivos compreende vários genes do metabolismo de lipídeos. Este resultado corrobora estudos anteriores apresentando a grande importância dos lipídeos no metabolismo energético de *M. tuberculosis in vivo* [38,39,103]. Outros genes importantes requeridos exclusivamente por *M. tuberculosis* são transportadores das superfamílias MFS e ABC, além do regulador transcricional denominado tetR *helix-turn-helix transcriptional repressor* Rv3050c (Tabela Suplementar 2, material online). Genes da família tetR têm sido implicados em vários processos biológicos relacionados a condições de estresse [104]. Devido ao papel biológico dos genes desta família para a sobrevivência do *M. tuberculosis* no hospedeiro, o gene Rv3050c poderia ser um regulador transcricional crítico atuando durante o processo de infecção. Genes, como o Rv3050c, que tem funções fundamentais no curso da infecção no hospedeiro poderiam ser indicados como potenciais alvos para o desenvolvimento de drogas.

A bactéria *B. pseudomallei* é outra bactéria de apelo para o desenvolvimento de novas terapias antimicrobianas. Esta bactéria é responsável pelo desenvolvimento da melioidose. Por isso, a determinação do conjunto de genes essenciais para essa bactéria permite

investigar as peculiaridades biológicas deste organismo do ponto de visto genético e funcional para a proposição de genes alvos para drogas. Da mesma forma como para *M. tuberculosis*, os NOGs essenciais exclusivos em *B. pseudomallei* foram identificados (Tabela Suplementar 2, material online). Dentre todos os NOGs anotados em *B. pseudomallei*, 105 são essenciais somente neste organismo. Fazem parte desta coleção genes como transportadores ABC e enzimas metabólicas, confirmando o cenário metabólico complexo deste organismo. Outros genes identificados são reguladores transcricionais das famílias GntR, LysR e AraC [105,106] que poderiam estar relacionados à resistência a antibióticos em *B. pseudomallei* e a sua capacidade para ocupar uma ampla variedade de ambientes, desde solo até o interior de células hospedeiras [107]. Na análise multivariada para avaliar a diversidade dos genes essenciais, *B. pseudomallei* distanciou-se de outra espécie filogeneticamente próxima, a *B. thailandensis* (Figura 8). As duas espécies, que divergiram há aproximadamente 47 milhões de anos [90], possuem genomas grandes, grande número de genes codificadores de proteínas e apresentam várias regiões de plasticidade genômica. Portanto, é provável que outros fatores expliquem este resultado [90,108]. Considerando a escassez de opções terapêuticas e a classificação da *B. pseudomallei* como potencial arma biológica em bioterrorismo pelo centro de controle de doenças (*Center for Disease Control and Prevention*) dos Estados Unidos, o grupo de genes essenciais identificado constitui uma fonte valiosa de possíveis genes candidatos para validação experimental no desenvolvimento de novas drogas.

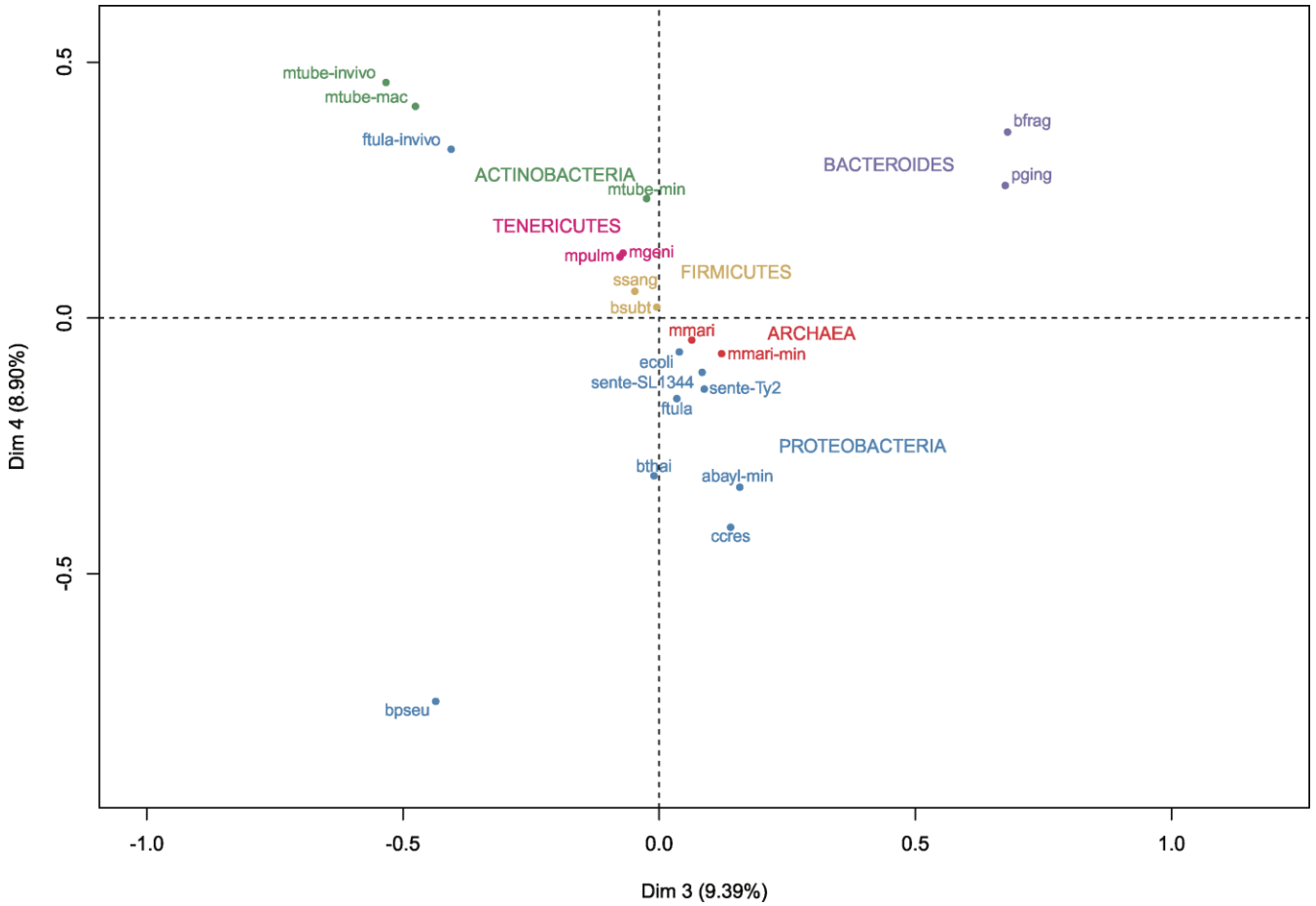


Figura 8 - MCA dos genes essenciais baseado na presença ou ausência do gene essencial em determinado NOG.
Legenda: Para abreviaturas vide Figura 5 e 7. Fonte: ilustração adaptada a partir da referência [4].

3.3. Genes essenciais universais

As anotações de NOG foram comparadas entre o organismo representante do domínio Archaea e os 15 organismos representantes do domínio Bacteria. Os NOGs essenciais em Archaea que não estão presentes no proteoma das bactérias foram considerados NOGs essenciais exclusivos de *M. maripaludis*. Por outro lado, NOGs anotados para genes essenciais em *M. maripaludis* que também estão presentes no proteoma de pelo menos uma bactéria, foram considerados NOGs compartilhados entre os dois domínios procarióticos. Dentre os 520 genes essenciais em meio rico para *M. maripaludis*, 10 genes apresentaram mais de uma anotação de NOG e por isso, foram desconsiderados da análise. Outros 194 genes foram exclusivos em arqueia (incluindo 26 genes sem anotação de NOG) e 316 genes apresentaram homólogos em bactéria. Dentre os 194 genes essenciais exclusivos, foi observado enriquecimento para as categorias funcionais “Replicação, recombinação e reparo” (L) e “Pobremente caracterizado” (S e R).

Três operons grandes (seis a oito genes) e inteiramente constituídos por genes essenciais exclusivos foram identificados em *M. maripaludis*. O operon do complexo enzimático *Tungstein-containing formylmethanofuran dehydrogenase* (DOOR operon ID 95836, Tabela Suplementar 3, material online) cataliza a desidrogenação do formilmetanofurano a metanofurano e monóxido de carbono (EC 1.2.99.5). Este complexo enzimático, presente em arqueias redutoras de sulfato e metanogênicas, catalisa os primeiros passos no processo da metanogênese. A metanogênese é responsável pela redução do dióxido de carbono a metano e pela fixação autotrófica de dióxido de carbono, sendo um processo crucial para o metabolismo [109]. Os outros dois operons são relacionados ao metabolismo e transporte de coenzimas. O operon da enzima *Tetrahydromethanopterin S-methyltransferase* (DOOR operon ID 95906) catalisa a formação de metil-coenzima M e tetraidrometanopterina a partir da coenzima M (EC 2.1.1.86) [110]. O outro operon da enzima *Methyl-coenzyme M reductase* (DOOR operon ID 95905), catalisa os passos finais da biossíntese de metano, reduzindo a metil-coenzima M e a coenzima B em metano (EC 2.8.4.1) [111]. A redução da emissão de dióxido de carbono tem se apresentado como um dos objetivos de órgãos ambientais para atenuar as consequências das mudanças climáticas globais. Durante a produção natural de metano por organismos metanogênicos ocorre a redução do dióxido de carbono e a síntese do metano. O metano pode ser utilizado como uma fonte de energia renovável na indústria. Desse modo, o cultivo em larga escala de

organismos metanogênicos e o entendimento das estruturas (operons e genes) envolvidas no processo de formação de metano têm grande potencial biotecnológico. Arqueias metanogênicas são, por exemplo, extremamente importantes na decomposição anaeróbica do esgoto; portanto, células otimizadas para a degradação de dióxido de carbono e produção de metano poderiam ser usadas no manejo do esgoto industrial e como uma fonte de energia renovável.

Os grupos de genes essenciais dos 16 organismos foram avaliados comparativamente com o objetivo de identificar os genes considerados essenciais em todos os organismos. Dezenove genes foram considerados essenciais nos 16 organismos testados *in vitro* (Tabela 4). A maioria dos genes pertence à categoria funcional “Armazenamento e processamento da informação” (J). Esta categoria compreende seis aminoacil-tRNA sintetases e oito proteínas ribossomais. Outras categorias presentes foram “Transcrição” (K), “Replicação, recombinação e reparo” (L), “Controle do ciclo e divisão celular” (D), “Tráfego intracelular” (U) e “Metabolismo e transporte de nucleotídeos” (F). A proteína SecY é a principal subunidade transmembrana do sistema de secreção do tipo II e a proteína Prs, converte a ribose 5-fosfato em fosforibosil pirofosfato (EC 2.7.6.1), que é essencial para o metabolismo de purinas. Estas observações estão de acordo com o fato de que tais genes estão envolvidos em processos biológicos centrais em todos os organismos. Entretanto, mesmo para os processos básicos dos organismos, têm sido descritos grupos de genes não ortólogos ou de genes distantemente relacionados realizando tais funções nas comparações entre Bacteria e Archaea [102,112,113]. Esse fenômeno explica o reduzido número de genes universais, especialmente quando apenas genes essenciais determinados experimentalmente são considerados. Estes achados demonstram a existência de um importante *core* de genes essenciais experimentalmente determinados que são compartilhados entre Bacteria e Archaea e que poderiam ter sido também essenciais no último ancestral comum entre esses grupos. À medida que houver informação experimental de essencialidade para outras bactérias e, principalmente, arqueias, poderemos avaliar com maior robustez este *core* de genes essenciais universais.

Tabela 4 – COG/NOGs universalmente conservados. Somente genes essenciais experimentalmente determinados foram considerados.

Armazenamento e Processamento da Informação		
Tradução, biogênese e estrutura ribossomal (J)	COG0018	Arginil-tRNA sintetase
	COG0008	Glutamil- e Glutaminil-tRNA sintetase
	COG0124	Histidil-tRNA sintetase
	COG0495	Leucil-tRNA sintetase
	COG0442	Prolil-tRNA sintetase
	COG0172	Seril-tRNA sintetase
	COG0090	Proteína ribossomal L2
	COG0087	Proteína ribossomal L3
	COG0088	Proteína ribossomal L4
	COG0097	Proteína ribossomal L6P/L9E
	COG0102	Proteína ribossomal L13
	COG0092	Proteína ribossomal S3
	COG0522	Proteína ribossomal S4
	COG0098	Proteína ribossomal S5
Transcrição (K)	COG0202	Subunidade alfa (40 kD) da RNA polimerase-DNA dependente
Replicação, Recombinação e Reparo (L)	COG0592	Subunidade beta da DNA polimerase III
Processos Celulares e Sinalização		
Controle do Ciclo Celular, Divisão Celular e Segregação do Cromossomo (D)	COG0037	Predita ATPase da superfamília PP-loop implicada no controle do ciclo celular
Tráfico Intracelular, Secreção e Transporte Vesicular (U)	COG0201	Preproteína translocase subunidade SecY
	COG0552	Partícula de reconhecimento de sinal GTPase (proteína FtsY)**
Metabolismo		
Transporte e Metabolismo de Nucleotídeos (F)	COG0462	Fosforibosilpirofosfato sintetase

*Legenda: ** Este COG foi identificado como essencial conservado apenas quando comparando com M. maripaludis S2 no meio mínimo. Fonte: tabela adaptada a partir da referência [4].*

3.4. Famílias gênicas

As famílias multigênicas podem ser de grande valor adaptativo na evolução de novas funções e para fornecer *backups* bioquímicos [114-116]. A existência de uma forte correlação entre o número de genes parálogos (famílias multigênicas) e o número de genes codificadores de proteínas é descrita na literatura e foi confirmada para as espécies analisadas neste estudo (Figura 9). À medida que o tamanho do genoma aumenta, o número de famílias de proteínas aumenta, assim como o número de genes nessas famílias, proporcionando maior diversidade e também maior redundância de funções nestes organismos [117]. Entretanto, como visto anteriormente, o número de genes essenciais não apresenta correlação com o número de genes codificadores de proteínas (Figura 5). Esse fenômeno está relacionado à técnica de identificação dos genes essenciais e à redundância de funções, como mencionado anteriormente, sugerindo que genes parálogos possam compensar a função do gene perdido.

Neste trabalho, foi avaliada a correlação entre essencialidade e a presença de genes parálogos no genoma. Uma correlação negativa forte ($P < 10^{-5}$) foi encontrada entre a presença de genes parálogos no genoma e a presença de genes essenciais nas famílias gênicas para a maioria dos organismos testados *in vitro* (13/17) (Figura 9, Anexo 2). Isso significa que, sob condições ideais, genes essenciais predominam em famílias monogênicas. Por um lado, esse efeito poderia ser relacionado à técnica e à interpretação utilizadas para os ensaios de essencialidade. Os experimentos se baseiam na interrupção de genes individuais, permitindo que outro gene (ortólogo ou não) possa compensar o desequilíbrio causado e comprometer a determinação do gene como essencial, mesmo quando de fato, a função seria essencial. É possível que futuros trabalhos que venham a utilizar combinações de genes interrompidos possam revelar uma outra tendência. Por outro lado, o fato de o gene exercer uma função fundamental para a célula e não apresentar um *backup* funcional homólogo (ou seja, ser único no genoma), o torna um gene essencial legítimo de família monogênica diante da letalidade celular.

Quando a mesma análise foi realizada para estudos *in vivo*, verificou-se que os genes requeridos são predominantemente de famílias multigênicas (Figura 9). Três explicações são possíveis para esse resultado. A primeira está relacionada a uma limitação da técnica durante a obtenção dos mutantes testados. Após o ensaio de transposição, os mutantes são primeiramente crescidos em meio de cultura rico. Apenas os mutantes que sobrevivem são

utilizados no sistema *in vivo* (hospedeiro ou células). Nessa situação, a grande maioria dos mutantes para genes essenciais *in vitro* são eliminados e não são testados na análise subsequente. Neste sentido, os ensaios *in vivo* perdem um considerável número de genes provenientes de famílias monogênicas. A segunda explicação está relacionada à falta de saturação dos ensaios *in vivo*. Isso significa que uma parte dos genes provavelmente não foi testada durante o ensaio de transposição. Essa limitação pode afetar a determinação de verdadeiros genes requeridos que poderiam pertencer a famílias monogênicas. Por fim, a terceira explicação está relacionada às características biológicas destes organismos diante de uma condição não ideal para o crescimento. Em um sistema *in vivo*, o organismo enfrenta dificuldades de crescimento e sobrevivência, por exemplo em decorrência de estresse, de estabelecimento da infecção, de evasão do sistema imune. Muitos genes relacionados a estes eventos provêm de famílias multigênicas conhecidas pela sua plasticidade e que são selecionadas para sequências variantes diante de situações adversas. Isso poderia explicar o enriquecimento dos genes requeridos *in vivo* para famílias multigênicas.

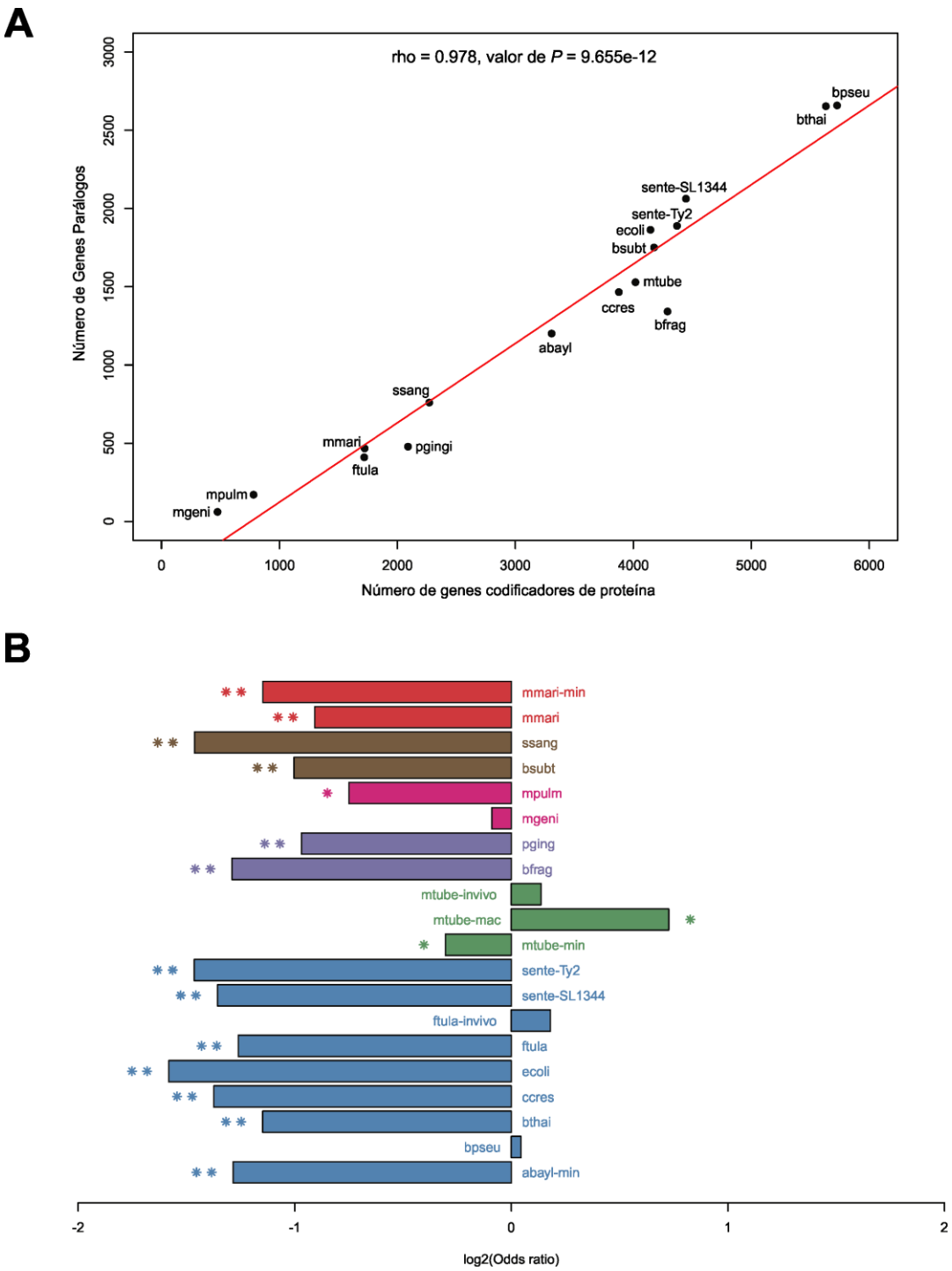


Figura 9 - Associação entre o número de genes codificadores de proteínas e da essencialidade dos genes com a presença de homólogos. A) Número total de genes codificadores de proteínas versus genes em famílias multigênicas: genes com a mesma assinatura de COG/NOG em um genoma foram considerados como parte de uma família multigênica, B) Essencialidade versus a presença de um homólogo no genoma. Teste exato de Fisher foi realizado para avaliar o enriquecimento dos genes essenciais em famílias multigênicas. Barras com um ou dois asteriscos representam $P \leq 10^{-2}$ e $P \leq 10^{-5}$, respectivamente.

Legenda: Para abreviaturas vide Figura 5 e 7. Fonte: ilustração adaptada a partir da referência [4].

3.5. Genes essenciais horizontalmente transferidos

Eventos de transferência lateral de genes são disseminados entre procariotos, contribuindo para a composição dos genomas [118]. Devido à importância da transferência horizontal, avaliou-se se genes horizontalmente transferidos poderiam contribuir para os repertórios essenciais em procariotos. Baseado no banco de dados DARKHORSE [81], um reduzido número de genes essenciais provenientes de transferência horizontal (< 2%) foi identificado nos repertórios essenciais de cada organismo (Tabela 5). Esse resultado encontra-se de acordo com a literatura na qual genes essenciais, genes cópia única e aqueles que definem o *core* genoma têm sido demonstrados como resistentes ao processo de transferência lateral [53,119].

Como o cerne das análises nesta seção número 2 da tese foi baseado em conceitos de ortologia, a avaliação da transferência horizontal é de grande importância para removermos a possibilidade de esses dados enviesarem os resultados apresentados. Para garantir que não houve viés, após a identificação dos genes transferidos, todas as análises foram refeitas. A remoção dos genes transferidos não afetou a significância estatística de nenhuma das análises.

Tabela 5 - Genes essenciais e genes não essenciais putativamente envolvidos em transferência gênica lateral de acordo com o banco de dados DARKHORSE.

Organismo	CDS[1]	HGT[2]	ESS[3]	essHGT[4]	nessHGT[5]	probID[6]	%(essHGT/ess)	%(nessHGT/ness)	%(HGT/CDS)
Acinetobacter baylyi ADP1	3307	152	499	17	135	0	3.41%	4.81%	4.60%
Escherichia coli K-12	4145	43	296	0	42	1*	0.00%	1.09%	1.04%
Francisella novicida U112	1719	214	390	24	190	0	6.15%	14.30%	12.45%
Salmonella typhi Ty2	4370	63	358	0	63	0	0.00%	1.57%	1.44%
Salmonella typhimurium SL1344		NA		NA	NA				
Burkholderia pseudomallei K96243	5727	66	505	2	64	0	0.40%	1.23%	1.15%
Burkholderia thailandensis E264	5632	273	406	2	271	0	0.49%	5.19%	4.85%
Caulobacter crescentus NA1000	3877	147	480	4	143	0	0.83%	4.21%	3.79%
Bacteroides fragilis 638R		NA		NA	NA				
Porphyromonas gingivalis ATCC 33277	2090	144	463	17	127	0	3.67%	7.81%	6.89%
Mycobacterium tuberculosis H37Rv	4018	339	771	30	308	1*	3.89%	9.49%	8.44%
Bacillus subtilis 168	4176	749	271	14	730	5*	5.17%	18.69%	17.94%
Streptococcus sanguinis SK36	2270	229	218	1	228	0	0.46%	11.11%	10.09%
Mycoplasma genitalium G37	475	10	378	7	3	0	1.85%	3.09%	2.11%
Mycoplasma pulmonis UAB CTIP	780	49	310	15	33	1*	4.84%	7.02%	6.28%
Methanococcus maripaludis S2 rico	1722	137	520	10	127	0	1.92%	10.57%	7.96%
Methanococcus maripaludis S2 mínimo	1722	137	651	19	118	0	2.92%	11.02%	7.96%

*Legenda [1] CDS, coding sequences ou genes codificadores de proteínas; [2] HGT, horizontal gene transfer ou genes horizontalmente transferidos; [3] ESS, genes essenciais; [4] essHGT, genes essenciais horizontalmente transferidos; [5] nessHGT, genes não essenciais horizontalmente transferidos; [6] probID, genes com problemas de identificador; * registro removido do banco de proteínas do NCBI. Fonte: tabela adaptada a partir da referência [4].*

4. Os repertórios de genes essenciais *in vitro* são distintos dos repertórios de genes requeridos *in vivo*

A identificação de genes requeridos *in vivo* é de grande interesse não somente para a pesquisa biomédica, mas também do ponto de vista evolutivo, uma vez que estes genes são provavelmente requeridos na natureza. As triagens *in vivo* são, em sua maioria, baseadas na aptidão do mutante. Quando um mutante falha em crescer ou mostra uma contagem reduzida no pool final, é inferido que o gene interrompido é importante para a infecção e sobrevivência do microrganismo no hospedeiro [38-40]. Dentre os estudos selecionados, grupos de genes únicos foram encontrados como requeridos *in vivo* e como essenciais *in vitro* (Figura 10). Como discutido anteriormente, a sobreposição entre os grupos *in vivo* e *in vitro*, pode estar sub-representada porque os mutantes são crescidos *in vitro* antes da inoculação e muitos genes importantes *in vitro* provavelmente são fundamentais também para a sobrevivência *in vivo*. Entretanto, genes requeridos *in vivo,* mas não *in vitro,* são funcionalmente diversos (Figura 10), sendo candidatos interessantes para o desenvolvimento de novas drogas.

Ao longo da progressão da doença, o organismo (ou, no caso experimental, o mutante) deve colonizar, disseminar e persistir em um ambiente nutricionalmente instável e sob os mecanismos de ação do sistema imune. Por isso, muitos genes requeridos *in vivo* são relacionados às categorias de metabolismo (Figura 10). Além dessas categorias, aquelas relacionadas aos “Genes pobremente caracterizados” e “Sem função conhecida” foram as mais enriquecidas nesses experimentos *in vivo* (Figura 10). Uma vez que estes genes desempenham um papel importante, porém ainda desconhecido durante o processo de infecção, a arquitetura dos domínios das proteínas codificadas por esses genes foi analisada (Tabela Suplementar S6, material online). Cito alguns exemplos:

- Experimento *in vivo* para *F. novicida*:
    - gene FTN_0933: proteína hipotética, domínio Pfam encontrado PF05402, relacionada à biosíntese da coenzima pirrol-quinolinea-quinona;
    - gene FTN_1133: proteína hipotética, domínio Pfam encontrado PF01540, relacionada a adesinas de variação antigênica associadas à virulência;
    - gene FTN_1196: proteína hipotética, domínio Pfam encontrado PF02575,

relacionada ao reparo do DNA.

- Experimento *in vivo* para *M. tuberculosis*:
    - gene Rv0100: proteína hipotética, domínio Pfam encontrado PF00550, relacionada à síntese de àcidos graxos;
    - gene Rv0207c: proteína hipotética, domínio Pfam encontrado PF01936, relacionada ao processoma de RNA transportador e RNA ribossomal.

Estes resultados evidenciam a existência de um campo aberto para estudos fenotípicos e de genômica funcional que poderiam esclarecer o papel desses genes nas complexas interações patógeno-hospedeiro.

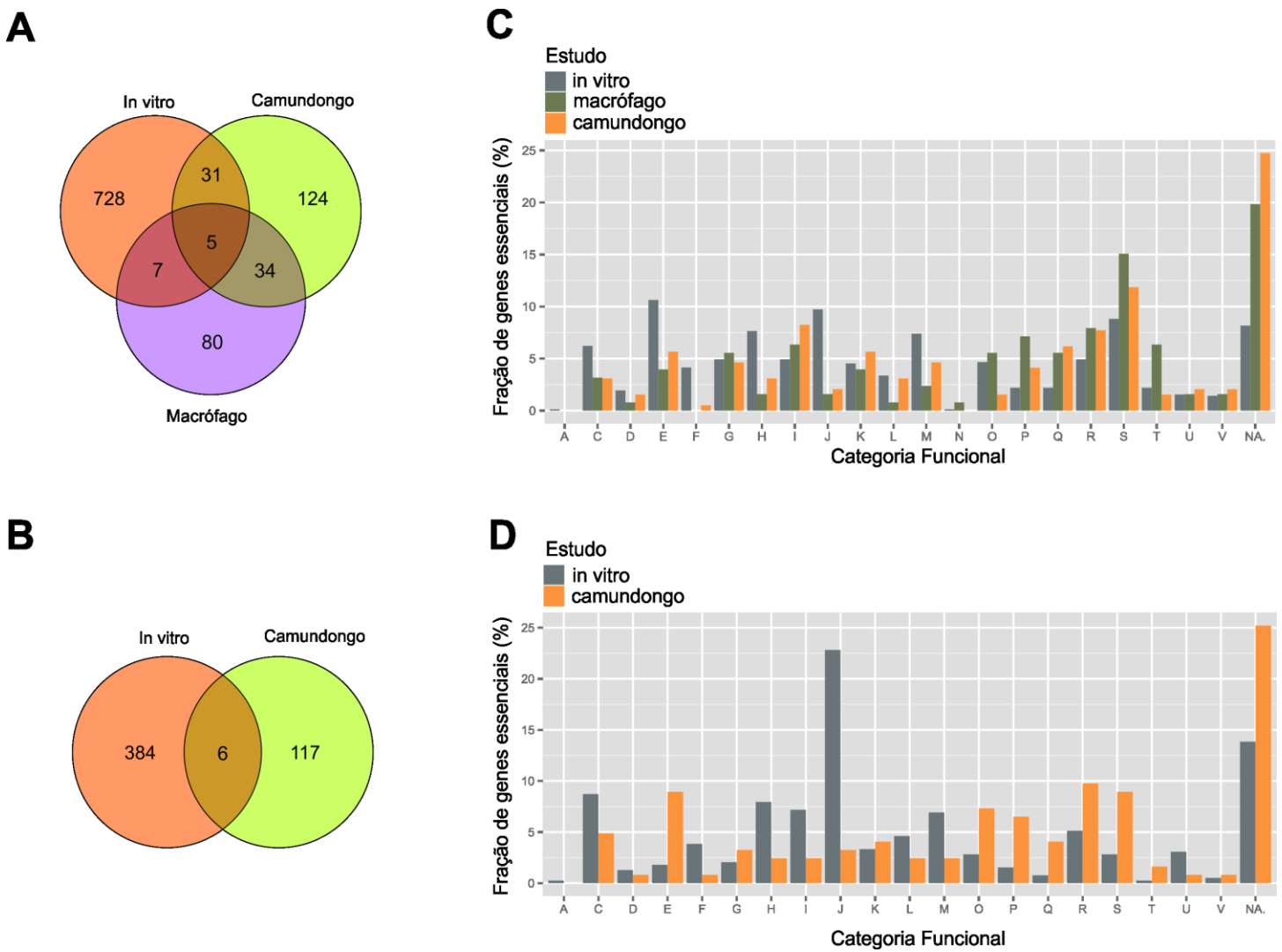


Figura 10 - Avaliação comparativa da essencialidade in vitro com o requerimento dos genes in vivo para os mesmos organismos. A) Diagrama de Venn mostrando o número de genes compartilhados e únicos para os experimentos in vivo (camundongo e macrófago) e in vitro de M. tuberculosis H37Rv, B) Diagrama de Venn mostrando o número de genes compartilhados e únicos para os experimentos in vivo (camundongo) e in vitro de Francisella tularensis novicida U112, C) Categorização funcional dos genes essenciais para os experimentos in vivo e in vitro de M. tuberculosis H37Rv, D) Categorização funcional dos genes essenciais para os experimentos in vivo e in vitro de F. tularensis novicida U112.

Fonte: ilustração adaptada a partir da referência [4].

5. Genes essenciais não são uniformemente distribuídos nos operons

Um estudo conduzido em 2003 demonstrou que os genes essenciais estão enriquecidos em operons em *E. coli* [52]. Embora várias listas de genes essenciais para outras bactérias tenham sido geradas ao longo dos anos, nenhum outro estudo procurou confirmar se a tendência observada para *E. coli* se mantinha nestas espécies, incluindo algumas taxonomicamente distantes. Visando cobrir esta lacuna, este projeto avaliou se a tendência mencionada acima seria uma característica geral para outras bactérias e para a arqueia.

A análise inicial revelou que os genes essenciais tendem a se apresentar como genes policistrônicos em todos os 16 procariotos estudados ($P \leq 0.05$, teste exato de Fisher) (Tabela 6). A influência dos genes que codificam proteínas ribossomais poderia ser questionada, uma vez que são genes essenciais *bonafide* e fortemente preservados nas mesmas estruturas de operons no genoma. Para responder a esse questionamento, a análise foi conduzida com a remoção dos genes para proteínas ribossomais, mostrando que a tendência observada é mantida (Anexo 3) e foi apoiada na comparação com 10 mil arquivos embaralhando os genes presentes em operons para cada um dos organismos testados (Anexo 4). Assim, a prevalência dos genes essenciais em operons parece ser uma característica antiga e disseminada na distribuição filogenética dos organismos. Essa característica poderia estar presente no último ancestral comum de Bactéria e Archaea. Por outro lado, a vantagem adaptativa de preservar os genes essenciais em operons poderia ser tão benéfica para a sobrevivência da célula, que poderia ter surgido independentemente em Bacteria e Archaea. Futuramente, a alternativa mais provável poderá ser indicada se um maior número de organismos presentes no domínio Archaea e outros grupos taxonômicos de Bacteria tiverem dados de essencialidade disponíveis e forem incorporados a este tipo de análise.

Uma vez que os genes essenciais tendem a se localizar em operons, a pergunta subsequentemente levantada foi: Genes essenciais apresentariam uma posição preferencial nos operons? Para responder a esta questão, a associação entre a ordem dos genes e a essencialidade foi avaliada. A ordem dos genes é normalmente preservada entre organismos proximamente relacionados. Essa conservação estrutural, entretanto, é progressivamente

perdida com o aumento da distância filogenética [10], exceto para uns poucos operons amplamente conservados, tais como os operons ribossomais [58-60]. Entretanto, mesmo estes operons ribossomais podem apresentar arranjos distintos entre os vários genomas bacterianos e de arqueias [60,61]. A análise de essencialidade no contexto da ordenação nos operons mostrou que os genes essenciais ocupam preferencialmente a primeira posição em operons contendo pelo menos um gene essencial (Tabela 7). Os genomas dos procariotos apresentam operons carregando um número variado de genes em sua estrutura. A análise inicial, entretanto, não considerou o número de genes nos operons. Desse modo, a frequência do tamanho dos operons (dado pelo número de genes em sua estrutura) foi avaliada em cada um dos genomas estudados. Os operons contendo dois ou três genes representam de 52.3 a 74.9% dos operons presentes nos organismos, conforme descrito previamente [49,120,121]. Devido a essa representatividade genômica, a análise estatística foi realizada somente para operons com pelo menos um gene essencial e de tamanhos dois e três. A análise de chi-quadrado, avaliando o número de genes essenciais observado em cada posição pelo número de genes essenciais esperado em cada posição do operon, confirmou o enriquecimento dos genes essenciais na primeira posição dos operons na maioria das espécies ($P < 0.01$), incluindo *M. maripaludis* (Tabela 8 e Tabela 9). Este resultado corrobora e refina aqueles obtidos por um estudo anterior que mostrou que os genes essenciais de *M. leprae* são enriquecidos na metade mais proximal do operon (metade 5'), enquanto pseudogenes tendem a estar presentes na região distal (metade 3') do operon [62].

Os procariotos apresentam aproximadamente 50% dos seus genes em operons [122]. E os genes essenciais são enriquecidos em operons em várias espécies (Tabela 7). Genes essenciais são mais conservados que os genes não essenciais ao longo da árvore filogenética (Figura 7), assim como, quanto à similaridade de sequência [46], provavelmente devido a maior taxa de expressão quando comparados com os genes não essenciais [69]. Além disso, genes essenciais tendem a ser *hubs* e a formar *cliques* (ou seja, sub-grafos completos) uns com os outros na rede de interação proteína-proteína [123,124]. Estas características podem contribuir para a tendência dos genes essenciais em se co-expressarem com outros genes, independentemente destes serem ou não essenciais sob condições similares. A organização dos genes em operons reduz a quantidade de regiões regulatórias necessárias à otimização a transcrição de genes co-regulados [53]. Além disso,

para o desenvolvimento de sistemas regulatórios complexos, o surgimento de operons é mais provável do que o de promotores independentes em genes distintos [53]. Esta observação é apoiada pela presença de regiões regulatórias mais complexas nos operons quando comparados a genes monocistrônicos [53].

Os genes policistrônicos são transcritos de acordo com a sua posição no operon (da extremidade 5' em direção a extremidade 3') e tem sido mostrada uma forte correlação entre o comprimento do operon, a ordem dos genes no operon e o nível de expressão gênica [125,126]. O nível expressão dos genes policistrônicos segue um padrão decrescente à medida que se aproxima do último gene do operon. Isso significa que genes na proximidade 5' apresentam maior nível de expressão comparados aos genes na extremidade 3' [125,126]. Baseado em dados de microarranjo e no índice de adaptação do códon [127], tem sido mostrado que os genes essenciais são mais expressos que os genes não essenciais [69,127]. Dessa forma, a presença dos genes essenciais na primeira posição dos operons (Tabela 8 e Tabela 9) apresenta implicações diretas nos seus maiores níveis de expressão em relação aos demais genes do operon [125]. Além disso, a presença do gene essencial na posição mais *upstream* (próxima ao promotor) no operon, preserva a chance de que o gene seja expresso se uma mutação prejudicar a transcrição em genes *downstream* (mais próximos ao códon de terminação). Considerando os resultados apresentados nesta tese e outros previamente publicados, sugere-se a hipótese de que as regiões regulatórias dos genes essenciais 5' poderiam guiar a regulação e, em última instância, a evolução do operon como um todo.

Tabela 6 - Distribuição dos genes essenciais e não essenciais de acordo com a organização do genoma (genes em operons versus genes monocistrônicos).

Distribuição dos Genes								
Organismo	**Genes Policistrônicos**			**Genes Monocistrônicos**			**Valor de *P***	**Odds ratio**
	Essencial	**Não Essencial**	**Total**	**Essencial**	**Não Essencial**	**Total**		
Porphyromonas gingivalis ATCC 33277	385	980	1365	78	647	725	1,75E-21	3,257073
Burkholderia pseudomallei K96243	403	3225	3628	102	1997	2099	6,87E-17	2,446183
Mycobacterium tuberculosis H37Rv	582	1968	2550	189	1279	1468	3,87E-15	2,000943
Escherichia coli K-12 MG1655	232	2388	2620	64	1461	1525	6,73E-09	2,219861
Salmonella enterica typhi Ty2	264	2318	2582	94	1694	1788	2,13E-09	2,052119
Burkholderia thailandensis S264	310	3211	3521	96	2015	2111	1,00E-09	2,026158
Acinetobacter baylyi ADP1	348	1601	1949	151	1208	1358	8,72E-08	1,737187
Bacillus subtilis	206	2298	2504	65	1607	1672	1,26E-08	2,215842
Francisella novicida U112	321	930	1251	69	399	468	8,12E-07	1,995186
Mycoplasma pulmonis UAB CTIP	247	298	545	63	172	235	1,05E-06	2,260616
Mycoplasma genitalium G37	343	75	418	35	22	57	7,07E-04	2,866586
Bacteroides fragilis 638R	407	2514	2921	140	1229	1369	6,91E-04	1,421081
Caulobacter crescentus NA1000	338	2090	2428	142	1307	1449	1,53E-04	1,488382
Salmonella enterica typhimurium SL1344	238	2472	2710	112	1624	1736	5,11E-03	1,395933
Methanococcus maripaludis S2 (rico)	351	715	1066	169	487	656	1,72E-03	1,414343
Streptococcus sanguinis SK36	159	1346	1505	59	706	765	2,90E-02	1,413323
Methanococcus maripaludis S2 (mínimo)	423	643	1066	228	428	656	4,09E-02	1,234764

Legenda: Espécies estão listadas de acordo com o nível de significância pelo teste exato de Fisher (do valor mais significativo para o menos significativo). Valor de P foi computado usando o teste exato de Fisher (P<0.05). Valores de Odds Ratio > 1 indicam enriquecimento em genes policistrônicos enquanto valores < 1 indicam enriquecimento em genes monocistrônicos. Fonte: tabela adaptada a partir da referência [4].

Tabela 7 - Relação entre o gene essencial e a posição no operon.

Organismo	Posição do gene essencial no operon													Ess-op[1]	Ess-genes[2]	Operons[3]
	1	2	3	4	5	6	7	8	9	10	11	13	14			
Acinetobacter baylyi ADP1	120 (65.2%)	50	7	3	3	1	0	0	0	0	0	0	0	184	348	643
Bacillus subtilis 168	59 (58.4%)	33	7	2	0	0	0	0	0	0	0	0	0	101	206	818
Bacteroides fragilis 638R	155 (69.1%)	42	19	4	1	1	2	0	0	0	0	0	0	224	407	956
Burkholderia pseudomallei K96243	126 (55.2%)	58	21	13	6	3	0	0	0	0	0	0	1	228	403	1146
Burkholderia thailandensis S264	97 (57.4%)	47	13	8	2	1	1	0	0	0	0	0	0	169	310	1155
Caulobacter crescentus NA1000	123 (67.2%)	38	14	3	5	0	0	0	0	0	0	0	0	183	338	844
Escherichia coli K-12 MG1655	67 (52.8%)	38	7	8	5	0	1	0	0	0	1	0	0	127	232	851
Francisella tularensis novicida U112	94 (59.9%)	42	12	7	2	0	0	0	0	0	0	0	0	157	321	373
Methanococcus maripaludis S2 (rico)	102 (62.2%)	44	13	3	1	1	0	0	0	0	0	0	0	164	351	362
Methanococcus maripaludis S2 (mínimo)	128 (62.7%)	56	12	5	3	0	0	0	0	0	0	0	0	204	423	362
Mycobacterium tuberculosis H37Rv	193 (59.2%)	92	24	10	4	1	0	2	0	0	0	0	0	326	582	895
Mycoplasma genitalium G37	69 (82.1%)	12	1	2	0	0	0	0	0	0	0	0	0	84	343	89
Mycoplasma pulmonis UAB CTIP	81 (74.3%)	21	5	2	0	0	0	0	0	0	0	0	0	109	247	175
Porphyromonas gingivalis ATCC 33277	144 (83.7%)	16	6	3	3	0	0	0	0	0	0	0	0	172	385	455
Salmonella enterica typhimurium SL1344	75 (57.7%)	37	10	3	3	2	0	0	0	0	0	0	0	130	238	881
Salmonella enterica typhi Ty2	76 (54.7%)	44	7	8	2	0	0	1	0	1	0	0	0	139	264	838
Streptococcus sanguinis SK36	50 (57.5%)	27	6	2	1	1	0	0	0	0	0	0	0	87	159	489

Legenda: [1] Ess-op, número de operons com pelo menos um gene essencial, [2] Ess-gene, número de genes essenciais policistrônicos, [3] Operons, número total de operons.
Fonte: tabela adaptada a partir da referência [4].

Tabela 8 - Posição dos genes essenciais nos operons com somente dois genes.

Organismo	Posição do gene essencial				X_2	DF	Valor de *P*
	1	2	Esperado	Total			
Acinetobacter baylyi ADP1	62	27	44,5	89	13,76	1	2,07E-04
Bacillus subtilis 168	30	16	23	46	4,26	1	3,90E-02
Bacteroides fragilis 638R	74	19	46,5	93	32,52	1	1,18E-08
Burkholderia pseudomallei K96243	64	28	46	92	14,08	1	1,75E-04
Burkholderia thailandensis S264	53	25	39	78	10,05	1	1,52E-03
Caulobacter crescentus NA1000	65	19	42	84	25,19	1	5,19E-07
Escherichia coli K-12 MG1655	38	17	27,5	55	8,01	1	4,63E-03
Francisella tularensis novicida U112	37	19	28	56	5,78	1	1,62E-02
Methanococcus maripaludis S2 (rico)	53	26	39,5	79	9,22	1	2,38E-03
Methanococcus maripaludis S2 (mínimo)	66	37	51,5	103	8,16	1	4,27E-03
Mycobacterium tuberculosis H37Rd	89	44	66,5	133	15,22	1	9,54E-05
Mycoplasma genitalium G37	19	5	12	24	8,16	1	4,27E-03
Mycoplasma pulmonis UAB CTIP	36	10	23	46	14,69	1	1,26E-04
Porphyromonas gingivalis ATCC 33277	64	11	37,5	75	37,45	1	9,36E-10
Salmonella enterica typhimurium SL1344	38	22	30	60	4,26	1	3,89E-02
Salmonella enterica typhi Ty2	42	27	34,5	69	3,26	1	7,10E-02
Streptococcus sanguinis SK36	19	15	17	34	0,47	1	4,93E-01

Fonte: tabela adaptada a partir da referência [4].

Tabela 9 - Posição dos genes essenciais em operons com somente três genes.

Organismo	Posição do primeiro gene essencial					X_2	DF	Valor de *P*
	1	2	3	Esperado	Total			
Acinetobacter baylyi ADP1	33	11	5	16,33	49	26,61	2	1,66E-06
Bacillus subtilis 168	12	4	6	7,33	22	4,72	2	9,41E-02
Bacteroides fragilis 638R	43	11	12	22	66	30,09	2	2,92E-07
Burkholderia pseudomallei K96243	22	8	9	13	39	9,38	2	9,17E-03
Burkholderia thailandensis S264	22	11	7	13,33	40	9,05	2	1,08E-02
Caulobacter crescentus NA1000	24	12	9	15	45	8,4	2	1,50E-02
Escherichia coli K-12 MG1655	8	11	2	7	21	6	2	4,98E-02
Francisella tularensis novicida U112	23	6	4	11	33	19,81	2	4,97E-05
Methanococcus maripaludis S2 (rico)	22	9	8	13	39	9,38	2	9,17E-03
Methanococcus maripaludis S2 (mínimo)	29	9	9	15,66	47	17,02	2	2,01E-04
Mycobacterium tuberculosis H37Rd	46	20	14	26,66	80	21,70	2	1,94E-05
Mycoplasma genitalium G37	18	2	0	6,66	20	29,20	2	4,56E-07
Mycoplasma pulmonis UAB CTIP	22	4	3	9,66	29	23,65	2	7,30E-06
Porphyromonas gingivalis ATCC 33277	36	3	3	14	42	51,85	2	5,49E-12
Salmonella enterica typhimurium SL1344	16	3	4	7,66	23	13,65	2	1,09E-03
Salmonella enterica typhi Ty2	12	4	3	6,33	19	7,68	2	2,15E-02
Streptococcus sanguinis SK36	16	4	0	6,66	20	20,80	2	3,04E-05

Fonte: tabela adaptada a partir da referência [4].

CONCLUSÕES

- Dentre os vários aspectos que afetam a determinação da essencialidade, nossa análise comparativa mostrou a influência tanto da complexidade biológica do organismo quanto da técnica utilizada;

- Uma mesma função pode ser realizada por genes não ortólogos entre os grupos taxonômicos distintos. Dessa forma, a essencialidade celular deve ser pensada como um conjunto de funções, que em consequência, resultará no conjunto de genes essenciais;

- Embora os genes essenciais tenham índices de persistência elevados, baseado em análise de ortólogos, apenas poucos genes apresentaram-se essenciais em todos os organismos in vitro. Esse fato pode estar relacionado à presença de genes não ortólogos até mesmo na maquinaria básica da célula;

- Os requerimentos funcionais no hospedeiro são distintos do crescimento em condições ideais. Funções executadas por genes em famílias gênicas são recrutadas, apresentando um repertório de genes in vivo como potenciais alvos de drogas para serem estudados;

- A presença dos genes essenciais nas primeiras posições dos operons poderia estar relacionada à proximidade com as regiões regulatórias, possivelmente também essenciais, e para as quais existe grande seleção purificadora a fim de evitar mutações;

- As características de essencialidade foram compartilhadas entre as 15 bactérias e a espécie de Archaea estudada. Esses resultados indicam a possibilidade de que estas características poderiam estar presentes no ancestral comum entre os domínios. Entretanto, não se pode descartar a possibilidade de que estas características tenham aparecido em eventos independentes entre os organismos.

CONSIDERAÇÕES FINAIS

Nesta tese é descrita a análise sistemática e comparativa dos genes essenciais experimentalmente determinados em vários organismos procariotos. As análises computacionais e experimentais para a determinação de genes essenciais são complementares. Essa afirmação torna-se evidente na análise de enriquecimento de funções entre os grupos de genes essenciais experimentais que seriam desconsideradas em análises exclusivamente computacionais. Além disso, foi demonstrado que os genes essenciais são tipicamente monogênicos e são mais conservados que os genes não essenciais ao longo da distribuição filogenética quando testados em condições ideais de crescimento. A persistência dos genes está relacionada às altas taxas de retenção gênica, que por sua vez, tem sido relacionada à essencialidade e à organização genômica [68]. A avaliação da persistência gênica tem se mostrado útil para a identificação de genes essenciais legítimos que não são detectados em condições de crescimento ideais (como exemplo, genes de reparo de DNA) [68]. Entretanto, muitos genes essenciais para o crescimento *in vivo* não são amplamente conservados e são frequentemente recrutados a partir de famílias multigênicas, sugerindo a existência de diferentes estratégias de sobrevivência que co-evoluíram com o respectivo hospedeiro. Este trabalho propõe novos alvos gênicos para o desenvolvimento de novas drogas e vacinas gênicas por meio da análise do padrão filético dos genes. Além do mais, genes essenciais não apenas predominam em operons, mas também tendem a ocupar a primeira posição nestas estruturas, reforçando a importância de suas regiões regulatórias em guiar a co-expressão de genes do operon. Destaca-se que a maioria das tendências encontradas para os genes essenciais são compartilhadas entre Bacteria e Archaea, sugerindo que estas características possam ter sido presentes no último ancestral comum entre estes domínios da vida. Alternativamente, é possível também que estas grandes tendências genômicas tenham evoluído independentemente, de forma convergente. A indústria biotecnológica, atuando no desenvolvimento de genomas bacterianos sintéticos, poderá se beneficiar de estudos computacionais e experimentais baseados nas características dos genes essenciais e persistentes, como o apresentado nesta tese, para o desenvolvimento de novas tecnologias.

REFERÊNCIAS BIBLIOGRÁFICAS

1. Gibson DG, Benders GA, Andrews-Pfannkoch C, Denisova EA, Baden-Tillson H, et al. (2008) Complete chemical synthesis, assembly, and cloning of a Mycoplasma genitalium genome. Science 319: 1215-1220.
2. Gibson DG, Glass JI, Lartigue C, Noskov VN, Chuang RY, et al. (2010) Creation of a bacterial cell controlled by a chemically synthesized genome. Science 329: 52-56.
3. Annaluru N, Muller H, Mitchell LA, Ramalingam S, Stracquadanio G, et al. (2014) Total synthesis of a functional designer eukaryotic chromosome. Science 344: 55-58.
4. Grazziotin AL, Vidal NM, Venancio TM (2015) Uncovering major genomic features of essential genes in Bacteria and a methanogenic Archaea. FEBS J 282: 3395-3411.
5. McCutcheon JP, Moran NA (2012) Extreme genome reduction in symbiotic bacteria. Nat Rev Microbiol 10: 13-26.
6. Martiny JB, Bohannan BJ, Brown JH, Colwell RK, Fuhrman JA, et al. (2006) Microbial biogeography: putting microorganisms on the map. Nat Rev Microbiol 4: 102-112.
7. Battistuzzi FU, Feijao A, Hedges SB (2004) A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land. BMC Evol Biol 4: 44.
8. Koonin EV (2009) Evolution of genome architecture. Int J Biochem Cell Biol 41: 298-306.
9. Omelchenko MV, Makarova KS, Wolf YI, Rogozin IB, Koonin EV (2003) Evolution of mosaic operons by horizontal gene transfer and gene displacement in situ. Genome Biol 4: R55.
10. Tamames J (2001) Evolution of gene order conservation in prokaryotes. Genome Biol 2: RESEARCH0020.
11. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, et al. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269: 496-512.
12. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, et al. (1995) The minimal gene complement of Mycoplasma genitalium. Science 270: 397-403.
13. Mushegian AR, Koonin EV (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proc Natl Acad Sci U S A 93: 10268-10273.
14. Koonin EV (2000) How many genes can make a cell: the minimal-gene-set concept. Annu Rev Genomics Hum Genet 1: 99-116.
15. Gil R, Silva FJ, Pereto J, Moya A (2004) Determination of the core of a minimal bacterial gene set. Microbiol Mol Biol Rev 68: 518-537, table of contents.
16. Koonin EV, Mushegian AR (1996) Complete genome sequences of cellular life forms: glimpses of theoretical evolutionary genomics. Curr Opin Genet Dev 6: 757-762.
17. Koonin EV, Mushegian AR, Galperin MY, Walker DR (1997) Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea. Mol Microbiol 25: 619-637.
18. Soll L, Berg P (1969) Recessive lethals: a new class of nonsense suppressors in Escherichia coli. Proc Natl Acad Sci U S A 63: 392-399.
19. Kessler DP, Englesberg E (1969) Arabinose-leucine deletion mutants of Escherichia coli B-r. J Bacteriol 98: 1159-1169.
20. Itaya M (1995) An estimation of minimal genome size required for life. FEBS Lett 362: 257-260.
21. Hutchison CA, Peterson SN, Gill SR, Cline RT, White O, et al. (1999) Global transposon mutagenesis and a minimal Mycoplasma genome. Science 286: 2165-2169.
22. Glass JI, Assad-Garcia N, Alperovich N, Yooseph S, Lewis MR, et al. (2006) Essential genes of a

minimal bacterium. Proc Natl Acad Sci U S A 103: 425-430.
23. Ji Y, Zhang B, Van SF, Horn, Warren P, et al. (2001) Identification of critical staphylococcal genes using conditional phenotypes generated by antisense RNA. Science 293: 2266-2269.
24. Akerley BJ, Rubin EJ, Novick VL, Amaya K, Judson N, et al. (2002) A genome-scale analysis for identification of genes required for growth or survival of Haemophilus influenzae. Proc Natl Acad Sci U S A 99: 966-971.
25. Chaudhuri RR, Allen AG, Owen PJ, Shalom G, Stone K, et al. (2009) Comprehensive identification of essential Staphylococcus aureus genes using Transposon-Mediated Differential Hybridisation (TMDH). BMC Genomics 10: 291.
26. Kobayashi K, Ehrlich SD, Albertini A, Amati G, Andersen KK, et al. (2003) Essential Bacillus subtilis genes. Proc Natl Acad Sci U S A 100: 4678-4683.
27. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, et al. (2006) Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol 2: 2006 0008.
28. de Berardinis V, Vallenet D, Castelli V, Besnard M, Pinet A, et al. (2008) A complete collection of single-gene deletion mutants of Acinetobacter baylyi ADP1. Mol Syst Biol 4: 174.
29. Xu P, Ge X, Chen L, Wang X, Dou Y, et al. (2011) Genome-wide essential gene identification in Streptococcus sanguinis. Sci Rep 1: 125.
30. Barquist L, Boinett CJ, Cain AK (2013) Approaches to querying bacterial genomes with transposon-insertion sequencing. RNA Biol 10: 1161-1169.
31. Luo H, Lin Y, Gao F, Zhang CT, Zhang R (2014) DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. Nucleic Acids Res 42: D574-580.
32. Chen WH, Minguez P, Lercher MJ, Bork P (2012) OGEE: an online gene essentiality database. Nucleic Acids Res 40: D901-906.
33. Acevedo-Rocha CG, Fang G, Schmidt M, Ussery DW, Danchin A (2013) From essential to persistent genes: a functional approach to constructing synthetic life. Trends Genet 29: 273-279.
34. Juhas M, Eberl L, Glass JI (2011) Essence of life: essential genes of minimal genomes. Trends Cell Biol 21: 562-568.
35. Klein BA, Tenorio EL, Lazinski DW, Camilli A, Duncan MJ, et al. (2012) Identification of essential genes of the periodontal pathogen Porphyromonas gingivalis. BMC Genomics 13: 578.
36. Gerdes SY, Scholle MD, Campbell JW, Balazsi G, Ravasz E, et al. (2003) Experimental determination and system level analysis of essential genes in Escherichia coli MG1655. J Bacteriol 185: 5673-5684.
37. Barquist L, Langridge GC, Turner DJ, Phan MD, Turner AK, et al. (2013) A comparison of dense transposon insertion libraries in the Salmonella serovars Typhi and Typhimurium. Nucleic Acids Res 41: 4549-4564.
38. Sassetti CM, Rubin EJ (2003) Genetic requirements for mycobacterial survival during infection. Proc Natl Acad Sci U S A 100: 12989-12994.
39. Rengarajan J, Bloom BR, Rubin EJ (2005) Genome-wide requirements for Mycobacterium tuberculosis adaptation and survival in macrophages. Proc Natl Acad Sci U S A 102: 8327-8332.
40. Kraemer PS, Mitchell A, Pelletier MR, Gallagher LA, Wasnick M, et al. (2009) Genome-wide screen in Francisella novicida for genes required for pulmonary and systemic infection in mice. Infect Immun 77: 232-244.
41. Weiss DS, Brotcke A, Henry T, Margolis JJ, Chan K, et al. (2007) In vivo negative selection screen identifies genes required for Francisella virulence. Proc Natl Acad Sci U S A 104: 6037-6042.
42. Chaudhuri RR, Peters SE, Pleasance SJ, Northen H, Willers C, et al. (2009) Comprehensive identification of Salmonella enterica serovar typhimurium genes required for infection of

BALB/c mice. PLoS Pathog 5: e1000529.
43. Umland TC, Schultz LW, MacDonald U, Beanan JM, Olson R, et al. (2012) In vivo-validated essential genes identified in Acinetobacter baumannii by using human ascites overlap poorly with essential genes detected on laboratory media. MBio 3.
44. Ge X, Kitten T, Chen Z, Lee SP, Munro CL, et al. (2008) Identification of Streptococcus sanguinis genes required for biofilm formation and examination of their role in endocarditis virulence. Infect Immun 76: 2551-2559.
45. Goodman AL, McNulty NP, Zhao Y, Leip D, Mitra RD, et al. (2009) Identifying genetic determinants needed to establish a human gut symbiont in its habitat. Cell Host Microbe 6: 279-289.
46. Jordan IK, Rogozin IB, Wolf YI, Koonin EV (2002) Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res 12: 962-968.
47. Luo H, Gao F, Lin Y (2015) Evolutionary conservation analysis between the essential and nonessential genes in bacterial genomes. Sci Rep 5: 13210.
48. Fang G, Rocha E, Danchin A (2005) How essential are nonessential genes? Mol Biol Evol 22: 2147-2156.
49. Conway T, Creecy JP, Maddox SM, Grissom JE, Conkle TL, et al. (2014) Unprecedented high-resolution view of bacterial operon architecture revealed by RNA sequencing. MBio 5: e01442-01414.
50. Okuda S, Kawashima S, Kobayashi K, Ogasawara N, Kanehisa M, et al. (2007) Characterization of relationships between transcriptional units and operon structures in Bacillus subtilis and Escherichia coli. BMC Genomics 8: 48.
51. Lawrence JG (1997) Selfish operons and speciation by gene transfer. Trends Microbiol 5: 355-359.
52. Pal C, Hurst LD (2004) Evidence against the selfish operon theory. Trends Genet 20: 232-234.
53. Price MN, Huang KH, Arkin AP, Alm EJ (2005) Operon formation is driven by co-regulation and not by horizontal gene transfer. Genome Res 15: 809-819.
54. Ussery D, Larsen TS, Wilkes KT, Friis C, Worning P, et al. (2001) Genome organisation and chromatin structure in Escherichia coli. Biochimie 83: 201-212.
55. Papp B, Pal C, Hurst LD (2003) Dosage sensitivity and the evolution of gene families in yeast. Nature 424: 194-197.
56. Rain JC, Selig L, De Reuse H, Battaglia V, Reverdy C, et al. (2001) The protein-protein interaction map of Helicobacter pylori. Nature 409: 211-215.
57. Sabatti C, Rohlin L, Oh MK, Liao JC (2002) Co-expression pattern from DNA microarray experiments as a tool for operon prediction. Nucleic Acids Res 30: 2886-2893.
58. Dandekar T, Snel B, Huynen M, Bork P (1998) Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci 23: 324-328.
59. Lathe WC, 3rd, Snel B, Bork P (2000) Gene context conservation of a higher order than operons. Trends Biochem Sci 25: 474-479.
60. Wolf YI, Rogozin IB, Kondrashov AS, Koonin EV (2001) Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. Genome Res 11: 356-372.
61. Coenye T, Vandamme P (2005) Organisation of the S10, spc and alpha ribosomal protein gene clusters in prokaryotic genomes. FEMS Microbiol Lett 242: 117-126.
62. Muro EM, Mah N, Moreno-Hagelsieb G, Andrade-Navarro MA (2011) The pseudogenes of Mycobacterium leprae reveal the functional relevance of gene order within operons. Nucleic Acids Res 39: 1732-1738.
63. Rocha EP, Danchin A (2003) Gene essentiality determines chromosome organisation in bacteria. Nucleic Acids Res 31: 6570-6577.

64. Lin Y, Gao F, Zhang CT (2010) Functionality of essential genes drives gene strand-bias in bacterial genomes. Biochem Biophys Res Commun 396: 472-476.
65. McLean MJ, Wolfe KH, Devine KM (1998) Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. J Mol Evol 47: 691-696.
66. Rocha EP, Danchin A (2003) Essentiality, not expressiveness, drives gene-strand bias in bacteria. Nat Genet 34: 377-378.
67. Price MN, Alm EJ, Arkin AP (2005) Interruptions in gene expression drive highly expressed operons to the leading strand of DNA replication. Nucleic Acids Res 33: 3224-3234.
68. Fang G, Rocha EP, Danchin A (2008) Persistence drives gene clustering in bacterial genomes. BMC Genomics 9: 4.
69. Rocha EP, Danchin A (2004) An analysis of determinants of amino acids substitution rates in bacterial proteins. Mol Biol Evol 21: 108-116.
70. Sarmiento F, Mrazek J, Whitman WB (2013) Genome-scale analysis of gene function in the hydrogenotrophic methanogenic archaeon Methanococcus maripaludis. Proc Natl Acad Sci U S A 110: 4726-4731.
71. Langridge GC, Phan MD, Turner DJ, Perkins TT, Parts L, et al. (2009) Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants. Genome Res 19: 2308-2316.
72. Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, et al. (2015) GenBank. Nucleic Acids Res 43: D30-35.
73. Mao X, Ma Q, Zhou C, Chen X, Zhang H, et al. (2014) DOOR 2.0: presenting operons and their functions through dynamic and integrated views. Nucleic Acids Res 42: D654-659.
74. Brouwer RW, Kuipers OP, van Hijum SA (2008) The relative value of operon predictions. Brief Bioinform 9: 367-375.
75. Eddy SR (2009) A new generation of homology search tools based on probabilistic inference. Genome Inform 23: 205-211.
76. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, et al. (2014) Pfam: the protein families database. Nucleic Acids Res 42: D222-230.
77. Powell S, Forslund K, Szklarczyk D, Trachana K, Roth A, et al. (2014) eggNOG v4.0: nested orthology inference across 3686 organisms. Nucleic Acids Res 42: D231-239.
78. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402.
79. Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 278: 631-637.
80. S L, J J, F H (2008) FactoMineR: an R package for multivariate analysis. Journal of Statisitics Software 25: 18.
81. Podell S, Gaasterland T, Allen EE (2008) A database of phylogenetically atypical genes in archaeal and bacterial genomes, identified using the DarkHorse algorithm. BMC Bioinformatics 9: 419.
82. Giaever G, Chu AM, Ni L, Connelly C, Riles L, et al. (2002) Functional profiling of the Saccharomyces cerevisiae genome. Nature 418: 387-391.
83. Kim DU, Hayles J, Kim D, Wood V, Park HO, et al. (2010) Analysis of a genome-wide set of gene deletions in the fission yeast Schizosaccharomyces pombe. Nat Biotechnol 28: 617-623.
84. French CT, Lao P, Loraine AE, Matthews BT, Yu H, et al. (2008) Large-scale transposon mutagenesis of Mycoplasma pulmonis. Mol Microbiol 69: 67-76.
85. Gallagher LA, Ramage E, Jacobs MA, Kaul R, Brittnacher M, et al. (2007) A comprehensive transposon mutant library of Francisella novicida, a bioweapon surrogate. Proc Natl Acad Sci U S A 104: 1009-1014.
86. Christen B, Abeliuk E, Collier JM, Kalogeraki VS, Passarelli B, et al. (2011) The essential genome

of a bacterium. Mol Syst Biol 7: 528.
87. Griffin JE, Gawronski JD, Dejesus MA, Ioerger TR, Akerley BJ, et al. (2011) High-resolution phenotypic profiling defines genes essential for mycobacterial growth and cholesterol catabolism. PLoS Pathog 7: e1002251.
88. Veeranagouda Y, Husain F, Tenorio EL, Wexler HM (2014) Identification of genes required for the survival of B. fragilis using massive parallel sequencing of a saturated transposon mutant library. BMC Genomics 15: 429.
89. Baugh L, Gallagher LA, Patrapuvich R, Clifton MC, Gardberg AS, et al. (2013) Combining functional and structural genomics to sample the essential Burkholderia structome. PLoS One 8: e53851.
90. Moule MG, Hemsley CM, Seet Q, Guerra-Assuncao JA, Lim J, et al. (2014) Genome-wide saturation mutagenesis of Burkholderia pseudomallei K96243 predicts essential genes and novel targets for antimicrobial development. MBio 5: e00926-00913.
91. Swoboda JG, Campbell J, Meredith TC, Walker S (2010) Wall teichoic acid function, biosynthesis, and inhibition. Chembiochem 11: 35-45.
92. Boucher Y, Kamekura M, Doolittle WF (2004) Origins and evolution of isoprenoid lipid biosynthesis in archaea. Mol Microbiol 52: 515-527.
93. Koga Y, Kyuragi T, Nishihara M, Sone N (1998) Did archaeal and bacterial cells arise independently from noncellular precursors? A hypothesis stating that the advent of membrane phospholipid with enantiomeric glycerophosphate backbones caused the separation of the two lines of descent. J Mol Evol 46: 54-63.
94. Dibrova DV, Galperin MY, Mulkidjanian AY (2014) Phylogenomic reconstruction of archaeal fatty acid metabolism. Environ Microbiol 16: 907-918.
95. Perez-Gil J, Rodriguez-Concepcion M (2013) Metabolic plasticity for isoprenoid biosynthesis in bacteria. Biochem J 452: 19-25.
96. Wilding EI, Brown JR, Bryant AP, Chalker AF, Holmes DJ, et al. (2000) Identification, evolution, and essentiality of the mevalonate pathway for isopentenyl diphosphate biosynthesis in gram-positive cocci. J Bacteriol 182: 4319-4327.
97. Guberman JM, Fay A, Dworkin J, Wingreen NS, Gitai Z (2008) PSICIC: noise and asymmetry in bacterial division revealed by computational image analysis at sub-pixel resolution. PLoS Comput Biol 4: e1000233.
98. Haydon DJ, Stokes NR, Ure R, Galbraith G, Bennett JM, et al. (2008) An inhibitor of FtsZ with potent and selective anti-staphylococcal activity. Science 321: 1673-1675.
99. Vaughan S, Wickstead B, Gull K, Addinall SG (2004) Molecular evolution of FtsZ protein sequences encoded within the genomes of archaea, bacteria, and eukaryota. J Mol Evol 58: 19-29.
100. Makarova KS, Yutin N, Bell SD, Koonin EV (2010) Evolution of diverse cell division and vesicle formation systems in Archaea. Nat Rev Microbiol 8: 731-741.
101. Weiss DS (2004) Bacterial cell division and the septal ring. Mol Microbiol 54: 588-597.
102. Koonin EV, Mushegian AR, Bork P (1996) Non-orthologous gene displacement. Trends Genet 12: 334-336.
103. McKinney JD, Honer zu Bentrup K, Munoz-Elias EJ, Miczak A, Chen B, et al. (2000) Persistence of Mycobacterium tuberculosis in macrophages and mice requires the glyoxylate shunt enzyme isocitrate lyase. Nature 406: 735-738.
104. Alex R, Sozeri O, Meyer S, Dildrop R (1992) Determination of the DNA sequence recognized by the bHLH-zip domain of the N-Myc protein. Nucleic Acids Res 20: 2257-2263.
105. Gallegos MT, Schleif R, Bairoch A, Hofmann K, Ramos JL (1997) Arac/XylS family of transcriptional regulators. Microbiol Mol Biol Rev 61: 393-410.

106. Maddocks SE, Oyston PC (2008) Structure and function of the LysR-type transcriptional regulator (LTTR) family proteins. Microbiology 154: 3609-3623.
107. Thibault FM, Hernandez E, Vidal DR, Girardet M, Cavallo JD (2004) Antibiotic susceptibility of 65 isolates of Burkholderia pseudomallei and Burkholderia mallei to 35 antimicrobial agents. J Antimicrob Chemother 54: 1134-1138.
108. Holden MT, Titball RW, Peacock SJ, Cerdeno-Tarraga AM, Atkins T, et al. (2004) Genomic plasticity of the causative agent of melioidosis, Burkholderia pseudomallei. Proc Natl Acad Sci U S A 101: 14240-14245.
109. Vorholt JA, Thauer RK (1997) The active species of 'CO2' utilized by formylmethanofuran dehydrogenase from methanogenic Archaea. Eur J Biochem 248: 919-924.
110. Gartner P, Ecker A, Fischer R, Linder D, Fuchs G, et al. (1993) Purification and properties of N5-methyltetrahydromethanopterin:coenzyme M methyltransferase from Methanobacterium thermoautotrophicum. Eur J Biochem 213: 537-545.
111. Ermler U (2005) On the mechanism of methyl-coenzyme M reductase. Dalton Trans: 3451-3458.
112. Lecompte O, Ripp R, Thierry JC, Moras D, Poch O (2002) Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale. Nucleic Acids Res 30: 5382-5390.
113. Leipe DD, Aravind L, Koonin EV (1999) Did DNA replication evolve twice independently? Nucleic Acids Res 27: 3389-3401.
114. Gossani C, Bellieny-Rabelo D, Venancio TM (2014) Evolutionary analysis of multidrug resistance genes in fungi - impact of gene duplication and family conservation. FEBS J 281: 4967-4977.
115. Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, et al. (2003) Role of duplicate genes in genetic robustness against null mutations. Nature 421: 63-66.
116. Mendonca AG, Alves RJ, Pereira-Leal JB (2011) Loss of genetic redundancy in reductive genome evolution. PLoS Comput Biol 7: e1001082.
117. Pushker R, Mira A, Rodriguez-Valera F (2004) Comparative genomics of gene-family size in closely related bacteria. Genome Biol 5: R27.
118. Soucy SM, Huang J, Gogarten JP (2015) Horizontal gene transfer: building the web of life. Nat Rev Genet 16: 472-482.
119. Lerat E, Daubin V, Moran NA (2003) From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-Proteobacteria. PLoS Biol 1: E19.
120. Salgado H, Moreno-Hagelsieb G, Smith TF, Collado-Vides J (2000) Operons in Escherichia coli: genomic analyses and predictions. Proc Natl Acad Sci U S A 97: 6652-6657.
121. Zheng Y, Szustakowski JD, Fortnow L, Roberts RJ, Kasif S (2002) Computational identification of operons in microbial genomes. Genome Res 12: 1221-1230.
122. Price MN, Arkin AP, Alm EJ (2006) The life-cycle of operons. PLoS Genet 2: e96.
123. Ning K, Ng HK, Srihari S, Leong HW, Nesvizhskii AI (2010) Examination of the relationship between essential genes in PPI network and hub proteins in reverse nearest neighbor topology. BMC Bioinformatics 11: 505.
124. Lin CC, Juan HF, Hsiang JT, Hwang YC, Mori H, et al. (2009) Essential core of protein-protein interaction network in Escherichia coli. J Proteome Res 8: 1925-1931.
125. Lim HN, Lee Y, Hussein R (2011) Fundamental relationship between operon organization and gene expression. Proc Natl Acad Sci U S A 108: 10626-10631.
126. Guell M, van Noort V, Yus E, Chen WH, Leigh-Bell J, et al. (2009) Transcriptome complexity in a genome-reduced bacterium. Science 326: 1268-1271.
127. Dotsch A, Klawonn F, Jarek M, Scharfe M, Blocker H, et al. (2010) Evolutionary conservation of essential and highly expressed genes in Pseudomonas aeruginosa. BMC Genomics 11: 234.
128. Forsyth RA, Haselbeck RJ, Ohlsen KL, Yamamoto RT, Xu H, et al. (2002) A genome-wide

strategy for the identification of essential genes in Staphylococcus aureus. Mol Microbiol 43: 1387-1400.
129. Jacobs MA, Alwood A, Thaipisuttikul I, Spencer D, Haugen E, et al. (2003) Comprehensive transposon mutant library of Pseudomonas aeruginosa. Proc Natl Acad Sci U S A 100: 14339-14344.
130. Zhou Y, Minami T, Honda K, Omasa T, Ohtake H (2010) Systematic screening of Escherichia coli single-gene knockout mutants for improving recombinant whole-cell biocatalysts. Appl Microbiol Biotechnol 87: 647-655.
131. Widdowson PS, Masten T, Halaris AE (1991) Interactions between neuropeptide Y and alpha 2-adrenoceptors in selective rat brain regions. Peptides 12: 71-75.
132. Zhang YJ, Ioerger TR, Huttenhower C, Long JE, Sassetti CM, et al. (2012) Global assessment of genomic regions required for growth in Mycobacterium tuberculosis. PLoS Pathog 8: e1002946.

ANEXOS

ANEXO 1 – Perspectiva histórica dos trabalhos de identificação de genes essenciais

Ano Referência	Organismo	Objetivo do Estudo	Meio de Cultivo	Estratégia (denominação original em inglês)	Nível de Saturação	Número de genes essenciais	Transposon / plasmid
1995 [20]	*Bacillus subtilis* 168	Genoma mínimo	LB (rico)	Site-directed mutagenesis by transformation	Não	6 regiões deletadas (562 kb)	-
1999 [21]	*Mycoplasma genitalium* *Mycoplasma pneumoniae*	Genoma mínimo	SP_4 (rico) Hayflick (rico)	Global transposon mutagenesis + PCR + electrophoreses	Não	265-350 genes essenciais estimados	Tn4001tet
2002 [128]	*Staphylococcus aureus* RN450	Alvo para drogas	LB (rico)	Rapid shotgun antisense RNA	Não (250,000 mutantes)	658 candidatos essenciais	pEPSA5
2003 [26]	*Bacillus subtilis* 168	Genoma mínimo	LB (rico)	Gene-by-gene inactivation	-	192 experimentais + 79 preditos = 271 (88% genes com status)	pMUTIN2
2003 [129]	*Pseudomonas aeruginosa* PAO1	-	LB (rico)	Global transposon mutagenesis + sequencing (ABI 3700)	Próximo a saturação (30,100 mutantes) (5 inserções/gene)	300-400 essenciais estimados (lista final baseada em ortologia) (87% ness com status)	ISphoA/hah ISlacZ/hah
2006 [22]	*Mycoplasma genitalium* G37	Genoma mínimo	SP4 (rico)	Global transposon mutagenesis + Sanger sequencing (ABI 3730xl)	Próximo a saturação 6 inserções/gene	382 + 5 paralogos = 387	Tn4001tet
2006 [130]	*Escherichia coli* K-12	Estudo fenotípica para genes de função desconhecida	LB (rico)	In-frame single gene deletions	(7970 mutantes)	303 candidatos essenciais (96% genes com status)	pKD46
2007 [85]	*Francisella tularensis novicida* U112	Identificar fatores de virulência e alvos para drogas	TSAC (rico)	Sequence-defined transposon mutant library + sanger sequencing (ABI 3700)	Próximo a saturação (26,106 mutantes) (16,508 inserções únicas)	396 candidatos essenciais (84% genes com status)	T15, T17, T18, T20
2008 [28]	*Acinetobacter baylyi* ADP1	Estudo de metabolism	MASK (mínimo)	Single-gene-deletion	-	499 candidatos essenciais (93% genes com status)	pEVL386

2008 [84]	*Mycoplasma pulmonis* CT	Estudo de biologia básica e mecanismos de patologia básica	MA and MB	Global transposon mutagenesis + sequencing (ABI 3700)	Próximo a saturação (1,700 mutantes) (1,856 inserções únicas) (1 inserção/300 bp)	310 candidatos essenciais (82% genes com status)	Tn4001T pIVT-1
2008 [131]	*Pseudomonas aeruginosa* PA14		LB (rico)	Global transposon mutagenesis + sequencing (ABI 3700)	Próximo a saturação (38,976 mutantes) (4.3 inserções/gene)	335 essenciais propostos (lista final baseada em ortologia) (75% ness com status)	Mar2xT7/hah pMar2xT7 TnphoA/hah pRT731
2009 [45]	*Bacteroides thetaiotaomicron* VPI-5482	Validação de nova técnica, genes requeridos emu ma população microbiana	TYG (rico)	Insertion sequencing transposon mutagenesis + illumina sequencing (InSeq)	Próximo a saturação (11,000 mutantes)	325 candidatos essenciais (78% genes com status)	Mariner Himar1 pSAM_Bt
2009 [71]	*Salmonella enterica* typhi Ty2	Determinar genes essenciais em meio rico e em meio contendo bile	LB (rico)	Transposon directed insertion sequencing site (TRADIS)	Saturado (106 mutantes) (370,000 inserções únicas) (80 inserções/gene)	356 candidatos essenciais (99% genes com status)	Tn5
2011 [29]	*Streptococcus sanguinis* SK36	Estudo de metabolismo	BHI (rico)	Systematic gene replacement	(2.106 mutantes)	218 candidatos essenciais (99% genes com status)	Km cassette
2011 [87]	*Mycobacterium tuberculosis* H37Rv	Perfil fenotípico global	Mínimo definido	High-density transposon mutagenesis + massive parallel sequencing (illumina)	Saturado (106 mutantes) (370,000 inserções únicas) (80 inserções/gene) (44,350 inserções) (1 inserção/100 bp)	774 candidatos essenciais	Himar1-ITR
2011 [86]	*Caulobacter crescentus* NA1000	Entendimento da rede regulatória	PYE (rico)	Hyper-saturated transposon mutagenesis coupled with high-throughput sequencing (illumina)	Saturado (800,000 mutantes) (428,735 inserções únicas) (1 inserção/7.65 bp)	480 candidatos essenciais (95% genes com status)	Tn5 pXMCS2::Tn5Pxyl
2012 [35]	*Porphyromonas gingivalis* ATCC 33277	Validação de uma nova técnica, estudo de metabolismo	BAPHK (rico)	Global transposon mutagenesis + illumina sequencing (TnSeq)	Próximo a saturação (54,000 mutantes) (35,853 inserções únicas) (55 inserções/gene)	463 candidatos essenciais (97% genes com status)	Mariner Himar1 pSAM_Bt

2012 [132]	*Mycobacterium tuberculosis* H37Rv	Genômica funcional	7H10 (definido)	High-density transposon mutagenesis + massive parallel sequencing (illumina)	Não (100,000 mutantes) (36,488 inserções únicas) (64 inserções/sitio)	742 genes com unidades funcionais requeridas (96% genes com status)	Himar1-ITR
2013 [89]	*Burkholderia thailandensis* E264	Identificar genes essenciais, resolver estrutura e propriedades de potenciais alvos para drogas	TYE (rico)	Saturation level transposon mutagenesis and next generation sequencing (TnSeq)	Saturado (390,000 mutantes) (163,727 inserções) (>30 inserções/gene)	406 candidatos essenciais	T23 pLG99
2013 [70]	*Methanococcus maripaludis* S2	Fenótipo de genes com funções desconhecidas	McCm (rico)	Saturation mutagenesis technique + Illumina sequencing (TnSeq)	Provavelmente saturado (3.104 - 6.104 mutantes) (25 inserções/library 1) 16 inserções/library 2)	526 candidatos essenciais (78% genes com status)	Tn5
2013 [70]	*Methanococcus maripaludis* S2	Fenótipo de genes com funções desconhecidas	McN (mínimo)	Saturation mutagenesis technique + Illumina sequencing (TnSeq)	Provavelmente saturado (3.104 - 6.104 mutantes) (8 inserções/library1) (9 inserções/library2)	664 candidatos essenciais (~85% genes com status)	Tn5
2013 [37]	*Salmonella enterica* typhimurium SL1344	Investigar as diferenças entre Ty2 e SL1344	LB (rico)	Transposon directed insertion sequencing site (TRADIS)	Saturado (930,000 mutantes) (549,086 inserções únicas) (>100 inserções/gene)	353 candidatos essenciais (98% genes com status)	Tn5
2014 [90]	*Burkholderia pseudomallei* K96243	Identificar alvos para drogas	LB (rico)	Transposon directed insertion sequencing site (TRADIS)	Saturado (106 mutantes) (chr1=240,000 inserções únicas – 1 inserção/25 bp) (chr2=70,000 inserções únicas – 1 inserção/45 bp)	505 candidatos essenciais	Modified-Tn5Km2 pVT
2014 [88]	*Bacteroides fragilis* 638R	Identificar genes importantes para a sobrevivência e potenciais alvos para drogas	BHI (rico)	Transposon delivery vetor + illumina sequencing	Saturado (110,003 mutantes) (1764 genes – 1-5 hits) (1798 genes – 6-198 inserções)	550 candidatos essenciais (99% genes com status)	pSAM_Bt

ANEXO 2 – Estatística geral das famílias multigênicas.

Organismo	Número de Genes Essenciais	Número de Genes Essenciais Parálogos	Número de CDS	Número de Genes Parálogos	Número de Famílias Parálogas	Tamanho Médio da Família	% Genes em Famílias Parálogas	Tamanho Máximo da Família
Mycoplasma genitalium G37	378	48	475	61	27	2,3	12,8	4
Mycoplasma pulmonis UAB CTIP	310	52	780	171	63	2,7	21,9	6
Francisella tularensis novicida U112	390	52	1719	410	167	2,5	23,9	14
*Methanococcus maripaludi*s S2 (rico)	520	99	1722	467	180	2,6	27,1	7
Methanococcus maripaludis S2 (mínimo)	651	117	1722	467	180	2,6	27,1	8
Porphyromonas gingivalis ATCC 33277	463	68	2090	478	175	2,7	22,9	49
Streptococcus sanguinis SK36	218	36	2270	760	268	2,8	33,5	10
Acinetobacter baylyi ADP1 (mínimo)	499	104	3307	1201	411	2,9	36,3	13
*Caulobacter crescentu*s NA1000	480	99	3877	1466	466	3,1	37,8	14
*Mycobacterium tuberculosi*s H37Rv (mínimo)	771	263	4018	1529	470	3,3	38,1	19
Escherichia coli K12 MG1655	296	67	4145	1864	630	3	45	16
Bacillus subtilis 168	271	74	4176	1751	578	3	41,9	8
Bacteroides fragilis 638R	547	93	4290	1342	426	3,2	31,3	17
Salmonella enterica typhi Ty2	358	82	4370	1889	663	2,8	43,2	28
Salmonella enterica typhimurium SL1344	350	93	4446	2063	712	2,9	46,4	8
Burkholderia thailandensis E264	406	121	5632	2654	791	3,4	47,1	25
Burkholderia pseudomallei K96243	505	238	5727	2659	802	3,3	46,4	16

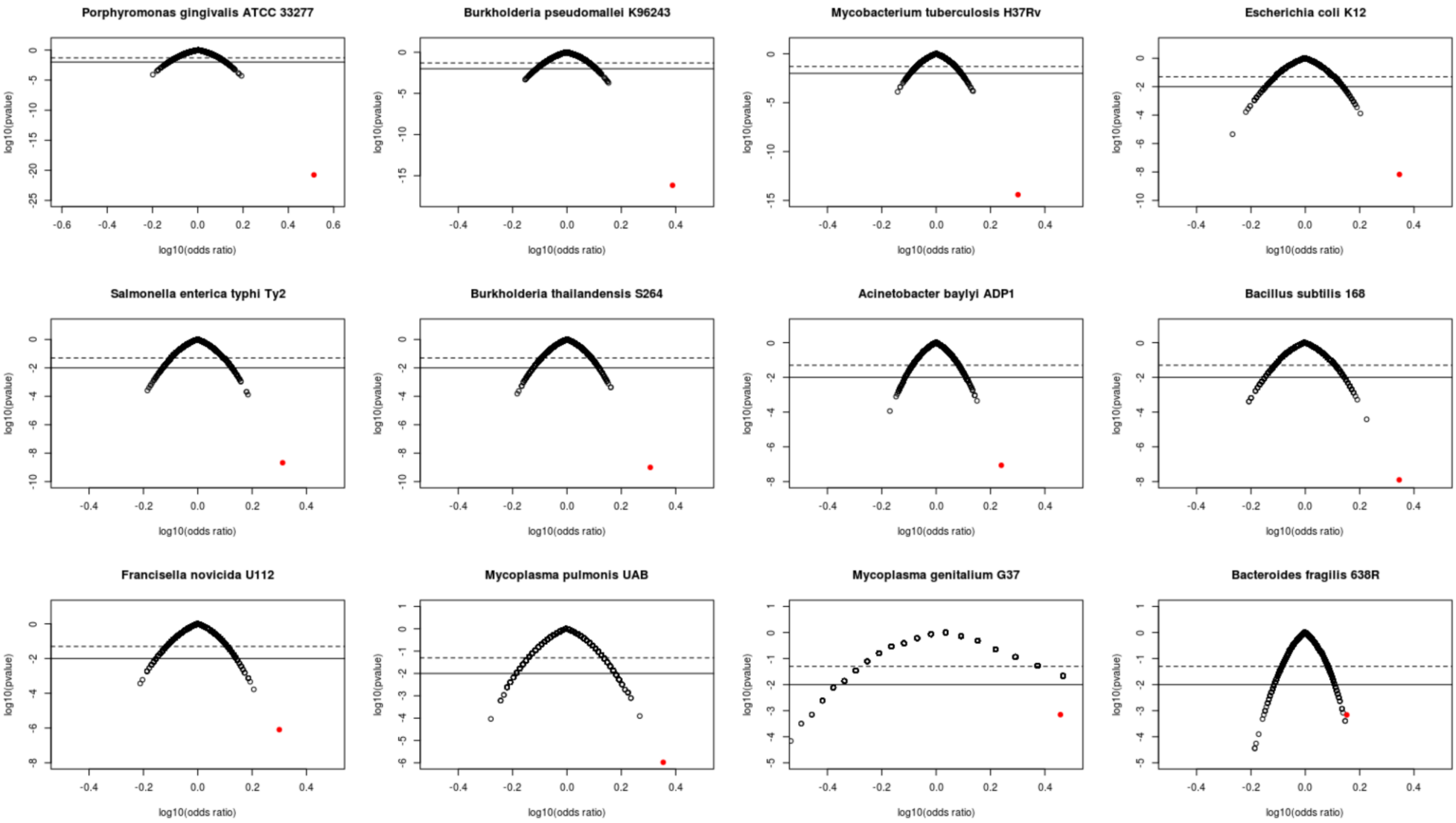
Legenda: Espécies estão listadas de acordo com o número de CDS (do menor para o maior repertório de genes codificadores de proteínas). Fonte: tabela adaptada a partir da referência [4].

ANEXO 3 – Distribuição dos genes codificadores de proteínas de acordo com a organização do genoma em genes policistrônicos e monocistrônicos, sem a participação dos genes de proteínas ribossomais.

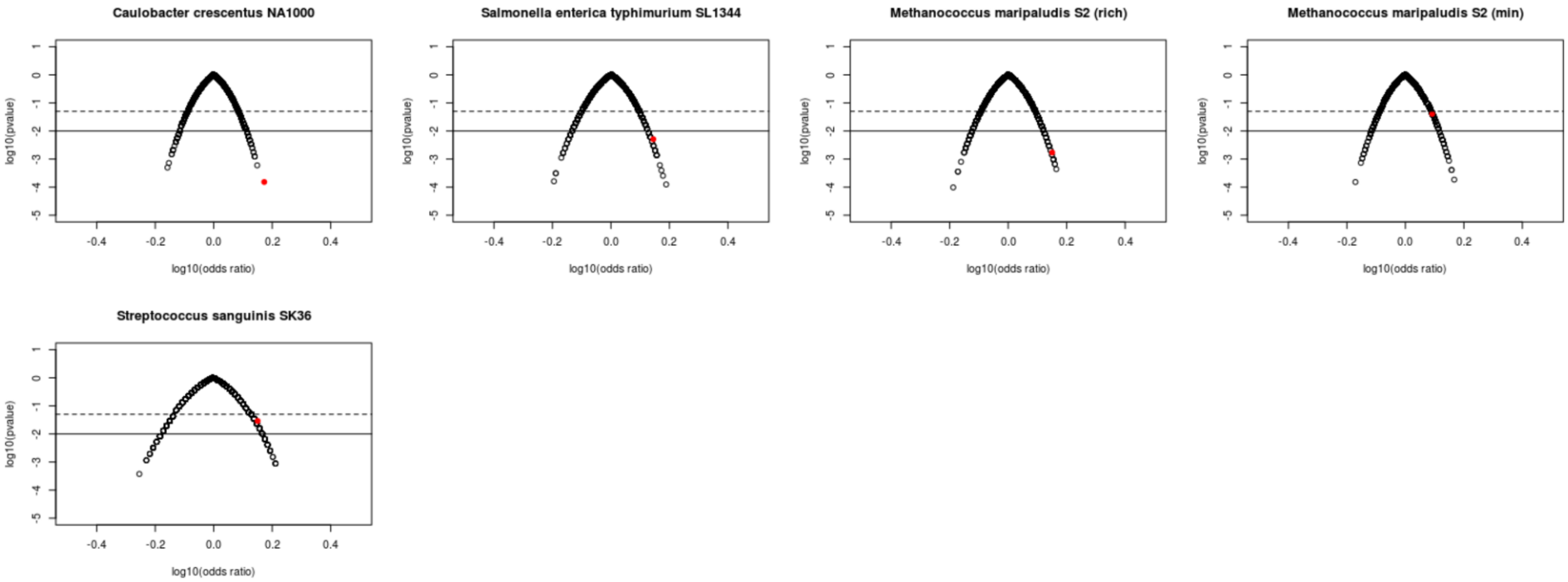
Organismo	Número de genes de proteínas ribossomais				Número de CDS sem os genes de proteínas ribossomais				Valor de *P*	Odds ratio
	Poli ess	Poli ness	Mono ess	Mono ness	Poli ess	Poli ness	Mono ess	Mono ness		
Porphyromonas gingivalis ATCC 33277	34	7	6	7	351	973	72	640	1,50E-19	3,21
Burkholderia pseudomallei K96243	14	32	0	10	389	3193	102	1987	1,22E-15	2,37
Mycobacterium tuberculosis H37Rv	24	27	0	6	558	1941	189	1273	1,02E-13	1,94
Burkholderia thailandensis S264	39	7	7	3	271	3204	89	2012	7,95E-08	1,91
Acinetobacter baylyi ADP1	41	5	7	2	307	1596	144	1205	8,99E-06	1,61
Escherichia coli K-12 MG1655	38	9	3	6	194	2378	61	1456	4,68E-06	1,95
Salmonella enterica typhi Ty2	43	6	5	2	221	2312	89	1692	2,68E-06	1,82
Bacillus subtilis 168	42	2	10	4	164	2296	55	1603	1,77E-06	2,08
Mycoplasma pulmonis UAB CTIP	43	2	5	2	204	296	58	170	6,13E-05	2,02
Francisella novicida U112	46	4	4	2	275	926	65	397	4,50E-05	1,81
Caulobacter crescentus NA1000	38	4	11	1	300	2086	131	1306	1,05E-03	1,43
Bacteroides fragilis 638R	38	3	9	3	369	2511	131	1226	3,04E-03	1,38
Mycoplasma genitalium G37	51	0	1	0	292	75	34	22	3,31E-03	2,51
Methanococcus maripaludis S2 (rico)	43	8	5	5	308	707	164	482	2,96E-02	1,28
Streptococcus sanguinis SK36	28	9	8	7	131	1337	51	699	8,65E-02	1,34
Salmonella enterica typhimurium SL1344	40	8	5	3	198	2464	107	1621	1,15E-01	1,22
Methanococcus maripaludis S2 (mínimo)	46	5	7	3	377	638	221	425	2,28E-01	1,14

Legenda: Espécies estão listadas de acordo com o nível de significância pelo teste exato de Fisher (do valor mais significativo para o menos significativo). Valor de P foi computado usando o teste exato de Fisher (P<0.05). Valores de Odds Ratio > 1 indicam enriquecimento em genes policistrônicos enquanto valores < 1 indicam enriquecimento em genes monocistrônicos. Fonte: tabela adaptada a partir da referência [4].

ANEXO 4 – Distribuição dos genes codificadores de proteínas de acordo com a organização do genoma em genes policistrônicos e monocistrônicos para os 10.000 arquivos simulados.



Illustração continua na página seguinte com a respectiva legenda.



Legenda: Os pontos pretos representam o valor de P e Odds ratio (dados transformados na escala logarítima) calculado para 10 mil arquivos simulados enquanto o ponto vermelho representa o valor real observado na espécie. Valor de P foi computado usando o teste exato de Fisher (P<0.05). Valores de Odds Ratio > 0.0 indicam enriquecimento em genes policistrônicos enquanto valores < 0.0 indicam enriquecimento em genes monocistrônicos. A linha pontilhada representa a significância de 0.05 enquanto a linha contínua representa a significância de 0.01.

ANEXO 5 - Resumo do trabalho de tese apresentado como pôster na seção de biologia computacional no *NIH Research Festival* em 16 de setembro de 2015.

Uncovering major genomic features of essential genes in Bacteria and a methanogenic Archaea

Wednesday, September 16, 2015 — Poster Session I

3:30 p.m. – 5:00 p.m. **FAES Terrace** **NLM** **COMPBIO-10**

Authors

- AL Grazziotin
- NM Vidal
- TM Venancio

Abstract

Understanding the organization and conservation of essential genes is critical for basic and applied biomedical research. Here we integrated a set of 17 high-resolution and genome-wide experimental in vitro essential gene studies and three in vivo required gene datasets, encompassing 15 bacteria and one Archaea. We assessed the overall features of essential genes in these two domains of life and demonstrated that most indispensable genes share important genomic architecture features, conservation and functional patterns among all species, possibly reflecting characteristics of an ancestral life form. Essential genes tend to be monogenic and are more conserved than nonessential genes. In contrast, genes particularly critical in vivo tend to be less conserved than those essential in vitro, suggesting that distinct strategies are deployed when the organism is stressed by the host immune system and unstable nutrient availability. Integration of evolutionary conservation and dispensability data allowed the identification of condition-specific novel gene targets with potential roles in the infection process of Mycobacterium tuberculosis and Burkholderia pseudomallei. Moreover, essential genes are not only preferentially located in operons, but are also biased toward the first position of the operons, supporting the influence of their regulatory regions in driving the transcription of whole operons. Finally, we demonstrated that the essential gene sets of Bacteria and an Archaea share important genomic features, indicating that high order properties of gene essentiality and genome architecture were probably present in the last universal common ancestor or evolved independently in the two prokaryotic domains of life.

Category: *Computational Biology*

ANEXO 6 – Trabalho de tese publicado na revista científica FEBS Journal.

the FEBS Journal

Uncovering major genomic features of essential genes in Bacteria and a methanogenic Archaea

Ana Laura Grazziotin[1,2,*], Newton M. Vidal[1,2,*] and Thiago M. Venancio[1]

1 Laboratório de Química e Função de Proteínas e Peptídeos, Centro de Biociências e Biotecnologia, Universidade Estadual do Norte Fluminense Darcy Ribeiro, Rio de Janeiro, Brazil
2 National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

Keywords
essential genes; genome evolution; genome organization; operons; prokaryotes; transposon mutagenesis

Correspondence
T. M. Venancio or A. L. Grazziotin, Laboratório de Química e Função de Proteínas e Peptídeos, Centro de Biociências e Biotecnologia, Universidade Estadual do Norte Fluminense Darcy Ribeiro, Av. Alberto Lamego 2000, P5/217 Campos dos Goytacazes, Rio de Janeiro, CEP 28013-602 – Brazil
Tel: +55 22 2748 6430
E-mails: thiago.venancio@gmail.com; analauragrazziotin@gmail.com

*These authors contributed equally to this work.

(Received 2 February 2015, revised 2 June 2015, accepted 15 June 2015)

doi:10.1111/febs.13350

Identification of essential genes is critical to understanding the physiology of a species, proposing novel drug targets and uncovering minimal gene sets required for life. Although essential gene sets of several organisms have been determined using large-scale mutagenesis techniques, systematic studies addressing their conservation, genomic context and functions remain scant. Here we integrate 17 essential gene sets from genome-wide *in vitro* screenings and three gene collections required for growth *in vivo*, encompassing 15 Bacteria and one Archaea. We refine and generalize important theories proposed using *Escherichia coli*. Essential genes are typically monogenic and more conserved than nonessential genes. Genes required *in vivo* are less conserved than those essential *in vitro*, suggesting that more divergent strategies are deployed when the organism is stressed by the host immune system and unstable nutrient availability. We identified essential analogous pathways that would probably be missed by orthology-based essentiality prediction strategies. For example, *Streptococcus sanguinis* carries horizontally transferred isoprenoid biosynthesis genes that are widespread in Archaea. Genes specifically essential in *Mycobacterium tuberculosis* and *Burkholderia pseudomallei* are reported as potential drug targets. Moreover, essential genes are not only preferentially located in operons, but also occupy the first position therein, supporting the influence of their regulatory regions in driving transcription of whole operons. Finally, these important genomic features are shared between Bacteria and at least one Archaea, suggesting that high order properties of gene essentiality and genome architecture were probably present in the last universal common ancestor or evolved independently in the prokaryotic domains.

Introduction

Bacteria and Archaea are widely diversified prokaryotic domains [1], adapted to a wide range of niches [2]. Prokaryotes evolved over billions of years and divergence of the major groups of Bacteria and of Archaea occurred between 2.5 and 3.2 and between 3.1 and 4.1 billion years, respectively [3]. Prokaryotic genomes have been carved by selection pressures, population size bottlenecks, mutation and recombination rates and mobile genetic elements [4], resulting in highly variable genome sizes and contents in different phylogenetic groups [5,6]. Genome sequencing efforts over the past two decades fueled the search for a universal set of

Abbreviations
CDS, coding sequence; COG, cluster of orthologous group; HGT, horizontal gene transfer; LUCA, last universal common ancestor; MCA, multiple correspondence analysis; NOG, non-supervised orthologous group.