MAGset: uma ferramenta para comparação de genomas recuperados de dados metagenômicos e sua aplicação para melhoria de suas montagens

Fabio Beltrame Sanchez

Dissertação apresentada
ao
Instituto de Matemática e Estatística
da
Universidade de São Paulo
para
obtenção do título
de
Mestre em Ciências

Programa Interunidades de Pós-graduação em Bioinformática

Orientador: Prof. Dr. João Carlos Setubal

São Paulo

2021

Fabio Beltrame Sanchez

MAGset: uma ferramenta para comparação de genomas recuperados de dados metagenômicos e sua aplicação para melhoria de suas montagens

Versão corrigida

Dissertação de Mestrado apresentada ao Programa Interunidades de Pós-Graduação em Bioinformática da Universidade de São Paulo para obtenção do título de Mestre em Ciências.

Área de Concentração: Bioinformática
Orientador: Prof. Dr. João Carlos Setubal

São Paulo

2021

Agradecimentos

À minha família, que criou a base para que hoje fosse possível ter a oportunidade de cursar a pós-graduação.

À minha esposa Silvia, pelo companheirismo, apoio e suporte nessa jornada. Agradeço também a sua família (agora minha também), meus sogros Marina e Silvio e cunhada Susan, sempre próximos e receptivos, dispostos a participar e ajudar em tudo.

À minha querida avó Mique, que com sua sede de conhecimento inspira a todos nós.

À minha mãe Ana, exemplo de positividade e energia infinita. Em memória do meu pai Reinaldo, que sempre acreditou nos estudos como o caminho para uma vida melhor. Aos meus irmãos, Hamilton e Heloisa, pela parceria de sempre.

Aos meus tios, tias e primas, por estarem sempre por perto, e tornar a vida em família mais completa e feliz.

Aos meus amigos, que desde o início me apoiaram nesse desafio (e em muitos outros) e acompanharam todos os passos, com destaque:

Ao André Delgado, entusiasta da área acadêmica entre os amigos, sempre dando dicas e acompanhando o processo.

Ao Denis Rosa, de colega de trabalho a grande amigo, exemplo de dedicação, determinação e foco. Ninguém segura.

Ao Eduardo Amuri, pelas muitas conversas e discussões sobre assuntos variados e complexos, parceiro e pau para toda obra. Nossos encontros nos cafés já geraram muito ciúmes na sociedade.

À Tamara Fonseca, que tem uma visão ímpar da vida e da sociedade, faz um belo contraponto e nos ajuda a entender outros pontos de vista. Agora com o Cleber, é uma dupla imbatível.

À turma do Setulab, pelos aprendizados e bons exemplos. Infelizmente não foi possível para mim conviver com todos o quanto eu gostaria, mas já valeu (e muito) a pena. Destaque ao Carlos Morais pelo suporte na infraestrutura, ao Robson Pontes por auxiliar nos testes e melhorias do software, e a todos que me ajudaram fornecendo os resultados/dados dos seus próprios trabalhos: Ana Carolina Soares, Livia Moura, Raquel Riyuzo e Suzana Guima.

Ao Professor Setubal, por toda sua disponibilidade, dedicação, paciência e envolvimento no projeto e trabalhos realizados.

Esse trabalho não é só meu, é uma conquista de todos nós.

Resumo

Sanchez, F. B. **MAGset: uma ferramenta para comparação de genomas recuperados de dados metagenômicos e sua aplicação para melhoria de suas montagens**. 2021. Dissertação (Mestrado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2021.

Um MAG (*metagenome-assembled genome*) é um genoma recuperado de dados metagenômicos e neste trabalho referem-se sempre a genomas de organismos procariotos. Após a obtenção de um MAG, diversas análises podem ser feitas para identificar similaridades e diferenças com os genomas já publicados da mesma espécie (quando a espécie é conhecida). Apresentamos MAGset, um software para comparar genomas e identificar especificidades em MAGs de espécies conhecidas. Essas especificidades podem ser regiões genômicas que existem somente no MAG e não existem nos genomas de referência, ou regiões que existem em um ou mais genomas de referência e não existem no MAG. Neste último caso, o módulo acessório MAGcheck permite verificar se as regiões não encontradas no MAG estão disponíveis nas amostras (*reads*) utilizadas na montagem do MAG, indicando um possível erro na montagem. Feita a comparação entre os genomas de interesse de forma automática pelo software, os seguintes resultados são apresentados ao usuário por meio de uma interface gráfica amigável: Matriz ANI comparando todos os genomas, pangenoma, anotações dos genes codificadores de proteína com os bancos CAZy e COG, regiões genômicas de interesse e resultado da validação do MAGcheck contra as amostras. Utilizando MAGset e MAGcheck, apresentamos os resultados de análises de 36 MAGs obtidos de diversas fontes. Os resultados obtidos com MAGcheck (obtivemos resultados em 34 MAGs) foram utilizados para realizar a remontagem dos MAGs originais, gerando melhorias na completude (24 dos 34 MAGs remontados) e no tamanho final (todos os MAGs remontados tiveram seu tamanho aumentado).

Palavras-chave: Metagenômica, MAG, montagem de genomas, genômica comparativa.

Abstract

Sanchez, F. B. **MAGset: a tool for comparing metagenome-assembled genomes and its application to improve their assembly**. 2021. Dissertação (Mestrado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2021.
A metagenome-assembled genome (MAG) is a genome reconstructed from metagenomic data, and in this work, MAGs refers to genomes of prokaryotic organisms. After obtaining a MAG, several analyses can be performed to identify similarities and differences with already published genomes from the same species (when the species is known). MAGset is a software to compare genomes and identify specificities in MAGs of known species. These specificities can be genomic regions that exist only in the MAG but not in reference genomes, or regions that exist in one or more reference genomes but do not exist in the MAG. In the latter case, the MAGcheck accessory module verifies whether the regions not found in the MAG are available in the samples (reads) used in the MAG assembly, indicating data missed by the assembler or binning program. Once the software has automatically compared the genomes of interest, the following results are presented to the user via a user-friendly graphical interface: ANI matrix comparing all genomes, pangenome information, annotations of protein-coding genes with CAZy and COG databases, genomic regions of interest, and the results of the MAGcheck validation against the samples. Using MAGset and MAGcheck, we present the results of analyses of 36 MAGs obtained from a variety of sources. The results obtained with MAGcheck (we obtained results for 34 MAGs) were used to perform the reassembly of the original MAGs, generating improvements in completeness (24 of the 34 MAGs reassembled) and in the final size (all reassembled MAGs increased in size).

Keywords: metagenomics, MAG, genome assembly, comparative genomics.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

ANI	Identidade média dos nucleotídeos (*Average Nucleotide Identity*)
pb	Pares de bases
COG	Agrupamento de proteínas em grupos ortólogos (*Clusters of Orthologous Groups of proteins*)
CSV	Valores separados por vírgula (*Comma separated values*)
FT	Referência ao artigo que contém dados de transferência fecal (*Fecal transfer*) [1]
GI	Referência ao artigo que contém dados intestinais de recém nascidos (*Gut infant*) [2]
HS	Referência ao artigo que contém dados de águas termais (*Hot springs*) [3]
MAG	Genoma montado a partir de dados metagenômicos (*Metagenome-assembled genome*)
Mpb	Milhões de pares de bases
PP	Referência ao artigo que contém dados de Fitoplânctons (*Phytoplankton*) [4]
RGI	Região genômica de interesse
SG	Referência ao artigo que contém dados de mudanças intestinais no tempo (*Shifts in the human gut*) [5]
SCG	Genes únicos universais (*Single Core Genes*)
SRA	Referência aos dados do estudo do Sistema Recifal Amazônico

Capítulo 1

Introdução

O estudo de microrganismos é importante para a compreensão do funcionamento de diversos ambientes, com aplicações em diferentes áreas, como a análise do solo para melhorias na produtividade de plantações, análise de amostras fecais para identificar doenças, entre outros.

Para compreender e estudar ambientes no nível de microrganismos, é possível obter amostras do ambiente de interesse e a partir destas amostras, extrair o DNA dos microrganismos ali presentes, em um processo chamado de extração de DNA metagenômico (de diversos genomas de uma só vez).

O DNA extraído pode ter como objetivo a análise de genes específicos para identificação de espécies de bactérias presentes (geralmente o gene RNA ribossomal 16S), e assim verificar quais espécies de bactérias estão contidas na amostras. Outra abordagem de análise, chamada de *shotgun*, tem como objetivo sequenciar todo o DNA possível, sem restringir a apenas um gene. O sequenciamento de todo DNA da amostra, apesar de mais complexo, tem grandes vantagens, como a possibilidade de se identificar novas funções nas bactérias do ambiente estudado, ao invés de apenas identificar quais bactérias existem ali [6]. No entanto, introduz desafios e dificuldades extras, já que nas tecnologias mais utilizadas para esse processo, o DNA extraído só pode ser obtido de forma fracionada (pequenas partes, chamadas de *reads*), e para realizar boa parte das análises, é necessário executar procedimentos para se obter a maior sequência contínua de DNA possível.

As sequências contínuas de DNA obtidas a partir da junção de diversos *reads* são chamadas de *contigs*. Após a obtenção de todos os *contigs* possíveis da amostra, é necessário identificar quais *contigs* possivelmente são do mesmo genoma, ou seja, juntar os *contigs* por similaridade, em um processo conhecido como *binning*. O resultado obtido ao final desse processo são arquivos com diversos *contigs*, e cada arquivo representa um possível genoma do ambiente analisado. Os genomas obtidos por esse processo são chamados de MAGs (*Metagenome-assembled genomes*, genomas obtidos de dados metagenômicos) [7].

Com o advento de novas tecnologias para obtenção de dados metagenômicos de forma mais fácil e barata, houve um aumento substancial da obtenção e publicação de MAGs nos trabalhos científicos. Diversas ferramentas e pipelines foram desenvolvidos para facilitar o processo de montagem dos MAGs, como o METAWrap [8], *pipeline* que integra diversas outras ferramentas necessárias para se obter MAGs a partir dos dados metagenômicos.

Ainda que façam essa tarefa com grande qualidade, o resultado não é perfeito, portanto devemos considerar os MAGs obtidos como possivelmente quiméricos e/ou incompletos, ou seja, podem conter partes de outros genomas que na montagem erroneamente foram considerados do seu genoma. Além disso, os MAGs são considerados como incompletos, pois várias partes podem ser perdidas no processo por diversos motivos: baixa cobertura

de algumas regiões no DNA extraído da amostra, *contigs* podem não ter sido agrupados corretamente no processo e terem sido descartados ou estarem em outros genomas, entre outras situações.

Uma vez obtido um ou mais MAGs de uma amostra, é um caminho natural verificar se já existem genomas publicados similares aos obtidos, ou seja, da mesma espécie. Ao identificar que o MAG é de uma espécie conhecida que já contém um genoma publicado, o próximo passo é compará-los para identificar o quanto são similares, quais características um genoma contém que o outro não contém, quais genes eles compartilham e quais são exclusivos, entre outros.

O panorama geral apresentado foi a motivação para o trabalho desta dissertação. Nesta dissertação apresentamos uma nova ferramenta, chamada de MAGset, que tem como objetivo auxiliar no processo de comparação entre MAGs e os genomas de referência da mesma espécie. Embora existam atualmente muitas ferramentas e plataformas que permitem comparação de genomas (como será visto no Capítulo 3.1), tanto quanto sabemos não existem ferramentas dedicadas a comparação de MAGs, que possibilitem a utilização dos resultados para melhorar a montagem/qualidade dos MAGs.

A principal função de MAGset é buscar regiões genômicas existentes em um genoma e não existente no outro, chamadas neste trabalho de Regiões Genômicas de Interesse (RGI). Essas regiões podem ser positivas (existentes no MAG e não existente nos genomas de referência) ou negativas (existentes em um ou mais genomas de referências e não existente no MAG).

As diferenças entre o MAG e o genoma de referência encontradas trazem uma nova questão: As partes faltantes no MAG e existentes no genoma de referência são erros de montagem? Essa questão é endereçada com o módulo MAGcheck, que realiza uma busca nas amostras utilizadas na montagem do MAG, tentando encontrar as regiões existentes no genoma de referência, mas não existente no MAG.

Caso a região seja encontrada nas amostras, surge mais uma questão: é possível usar os dados encontrados para melhorar o MAG original, gerando um genoma mais completo? Esta questão também foi endereçada e validada neste trabalho. Os resultados dessa remontagem que mostraremos na dissertação trouxeram em geral uma melhoria da montagem, com incremento no tamanho final dos MAGs, melhoria no valor de completude e consequentemente maior número de genes identificados no MAG remontado.

Esta dissertação está estruturada da seguinte forma:

- Capítulo 2: Conceitos utilizados neste trabalho;
- Capítulo 3: Trabalhos relacionados;
- Capítulo 4: Apresentação do software desenvolvido neste trabalho e a metodologia utilizada;
- Capítulo 5: Metodologia utilizada para a remontagem dos MAGs utilizando os resultados do software desenvolvido;
- Capítulo 6: Conjunto de dados utilizados para validar o software;
- Capítulo 7: Apresentação dos resultados obtidos utilizando o software e a metodologia para remontagem apresentada no Capítulo 5;
- Capítulo 8: Discussão dos resultados;
- Capítulo 9: Conclusão e trabalhos futuros;

- Apêndice A: Informações adicionais sobre as amostras metagenômicas e as ferramentas utilizadas para montagem dos MAGs pelos projetos e artigos analisados neste trabalho;
- Apêndice B: Resultados detalhados das comparações realizadas, descrevendo para cada MAG original, MAG remontado e referências: tamanho; completude; contaminação; quantidade de genes; RGIs; comparação entre o MAG original e o remontado.

Capítulo 2

Conceitos

2.1 Espécie

Uma espécie procariótica é considerada um grupo de cepas que se caracterizam por um certo grau de consistência fenotípica, mantendo 70% ligação DNA-DNA e mais de 97% da identidade da sequência genética do gene RNA ribossomal 16S [9]. Outro método utilizado para se definir se duas cepas são da mesma espécie é o ANI (definido em 2.15), método escolhido para ser utilizado neste trabalho, pela disponibilidade de diversos softwares que realizam esse tipo de comparação e por sua larga utilização em artigos e trabalhos de bioinformática.

2.2 Isolados e Cepas

Chamamos *isolado* a colônia homogênea de um microrganismo cultivado em laboratório, e cuja origem foi um processo de isolamento desse microrganismo a partir de um ambiente ou hospedeiro. Desta forma, o sequenciamento de um isolado é o processo de sequenciar o DNA obtido da colônia homogênea. Um isolado é um conceito que faz sentido na prática de cultivo laboratorial. Já o conceito de *cepa* (sinônimos: linhagem, estirpe) é um conceito taxonômico: refere-se à classificação de um organismo, num nível abaixo de espécie.

Assim sendo, isolados diferentes podem pertencer à mesma cepa. Ademais, uma cepa de um microrganismo pode ser reconhecida mesmo sem ter um isolado correspondente; é justamente o caso (em geral) de um MAG classificado como de uma espécie com genoma sequenciado, sem ser 100% idêntico a este.

2.3 Genomas de referência

Nesta dissertação, genomas de referência são aqueles com os quais iremos comparar os MAGs de interesse. Preferencialmente um genoma de referência deve ter sido sequenciado a partir de um isolado, mas em alguns casos utilizamos genomas de referência que são MAGs obtidos em projetos separados dos projetos que geraram os MAGs de interesse.

2.4 *Reads*

Reads são fragmentos genômicos obtidos após o processamento das amostras. Nos métodos mais comuns de processamento de amostras para extração de *reads*, chamados de

sequenciamento de segunda geração, os *reads* são curtos, geralmente fragmentos entre 75 e 400 pb (pares de bases), não sendo possível obter *reads* maiores que 1.000 pb por limitação da tecnologia. Já está disponível o sequenciamento de terceira geração, que permite a obtenção de *reads* longos (geralmente fragmentos entre 5.000 e 30.000 pb) [10].

Embora os *reads* longos sejam mais desejáveis, a tecnologia para obtê-los ainda tem uma taxa de erros por base maior que a tecnologia para *reads* curtos (entre 1% e 5% para *reads* longos e 0,1% para *reads* curtos) [10].

Alguns estudos em busca de melhor qualidade nos resultados obtidos estão usando conjuntos de *reads* longos e *reads* curtos da mesma amostra, para conseguir assim o melhor de cada tecnologia (sequências mais longas com maior qualidade) [11]. Neste trabalho foram utilizados apenas dados provenientes de *reads* curtos.

Os *reads* podem ser obtidos em dois formatos:

- *Single-end*: Apenas uma leitura é obtida a partir de um fragmento de DNA;
- *Paired-end*: A partir de um fragmento de DNA duas leituras são obtidas: uma da ponta $5'$ para a ponta $3'$ da fita $+$, e outra, da outra ponta, na fita oposta, também no sentido $5'$ para $3'$.

A utilização de *reads paired-end* pode permitir conexão de *contigs* (ver Seção 2.7).

2.5 Amostras

Neste trabalho definimos amostras como o conjunto de *reads* obtidos de determinado ambiente ou organismo. Caso a coleta seja de um local que contenha diversos organismos de diferentes espécies, como uma coleta do solo de uma região, essa amostra contém fragmentos genômicos de diferentes espécies, e a chamamos de amostra metagenômica.

2.6 *Contigs*

Ao realizar o processo de montar os *reads* que se sobrepõem para obter fragmentos maiores de sequências genômicas, cada fragmento maior é chamado de *contig*.

2.7 *Scaffolds*

Enquanto *contigs* são a união de diversos *reads*, *scaffolds* são compostos de dois ou mais *contigs*. Um *scaffold* é uma coleção de *contigs* ordenados. Essa ordenação procura reproduzir a ordem com que os *contigs* do *scaffold* aparecem no genoma. *Scaffolds* podem ser obtidos de diversos dados; um dos mais comuns é um par de *reads* que sejam pareados (*paired-end*). Se um dos membros do par está no *contig* A e o outro membro do par está num *contig* B, e se suas orientações são consistentes, podemos dizer que A é um *contig* adjacente a B no *scaffold* ao qual pertencem. Com este método é inclusive possível de se estimar o tamanho do sequenciamento faltante (*gap*) entre A e B, que será dado pelo tamanho do inserto do fragmento sequenciado que resultou nos *reads* pareados [12].

2.8 Montadores

Para se obter os *contigs/scaffolds* a partir dos *reads* de uma determinada amostra, são utilizados softwares chamados de montadores. Existem diversos montadores disponíveis, e alguns comumente utilizados são: Spades [13], MegaHIT [14] e IDBA-UD [15].

2.9 MAGs

Após obter os *contigs/scaffolds* utilizando montadores a partir de uma amostra metagenômica, é possível utilizar outro tipo de ferramenta para agrupá-los por suas características (em um processo conhecido como *binning*), e assim obter um grupo de *contigs/scaffolds* que são considerados como sendo do mesmo genoma. Os genomas obtidos a partir desse processo são chamados de MAGs (*Metagenome-assembled genomes* - genomas montados a partir de dados metagenômicos).

Assim como softwares de montagem, existem diversos softwares para se obter MAGs a partir dos *contigs/scaffolds*, e entre os mais utilizados temos o MetaBAT [16], CONCOCT [17] e MaxBin2 [18].

Embora associemos um MAG a um determinado genoma de um microrganismo, é possível que entre as sequências que definem um MAG se encontrem sequências que não pertencem a esse microrganismo; tais sequências são conhecidas como *contaminações*. O grau de contaminação de um MAG depende de vários fatores, entre eles o método de obtenção do MAG e a diversidade de microrganismos na amostra de onde vieram os *reads* que permitiram a montagem do MAG. Por outro lado, também não há garantia que a sequência de um MAG represente o genoma completo; de forma geral, pode-se afirmar que é rara a situação em que um MAG seja um genoma completo. Estes dois conceitos, de completude e de contaminação de um MAG, tem um papel importante nesta dissertação. O programa mais usado para se determinar completude e contaminação de MAGs (na verdade, de genomas em geral) chama-se CheckM [19].

2.10 Região genômica de interesse

Regiões genômicas de interesse (RGIs) são regiões existentes em um MAG e não existentes em nenhum genoma de referência para esse MAG (RGI positiva); ou então existente em pelo menos uma referência e não existente no MAG correspondente (RGI negativa).

A Figura 2.1 ilustra o conceito de RGI.

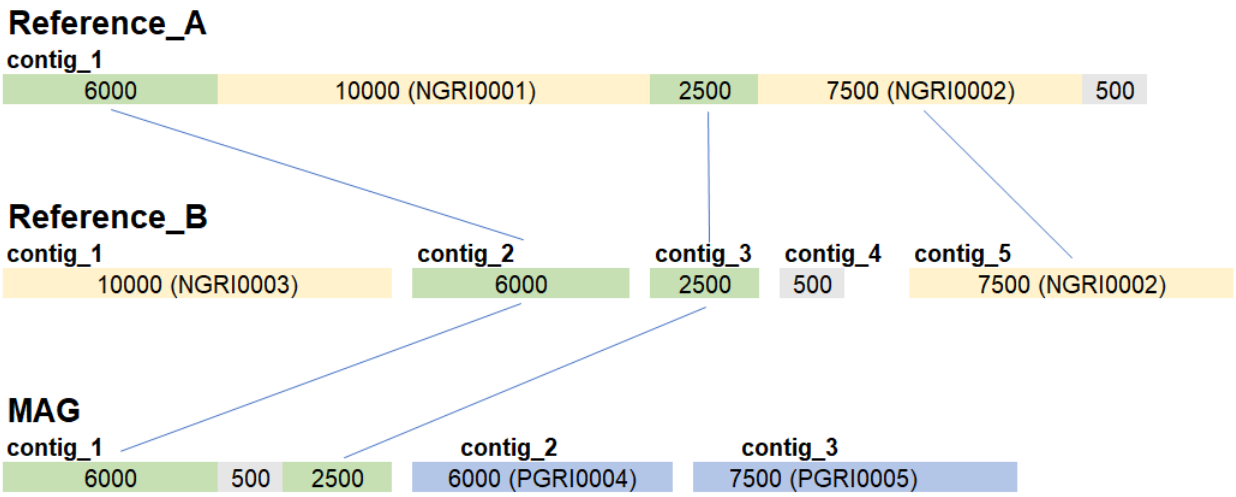


Figura 2.1: *Identificação de RGIs em MAGs e referências*
Exemplo de comparação de um MAG com duas referências. O valor dentro de cada retângulo representa o tamanho da região. Regiões em verde são as regiões similares em todos os genomas, ou seja, não são RGIs. As regiões amarelas são RGIs negativas (NGRI0001, NGRI0002) e as regiões azuis são RGIs positivas (PRGI0003 e PRGI0004). A RGI0002 foi encontrada em dois genomas de referência, portanto as duas regiões contêm o mesmo código identificador. Regiões em cinza são exclusivas, mas foram descartadas pois não cumpriram o requisito de tamanho mínimo.

Neste ponto da dissertação deixamos a especificação do tamanho mínimo de RGIs propositalmente vago. O tamanho mínimo de uma RGI será definido na apresentação do software que é tema deste trabalho. Mas de uma forma geral, vamos supor que RGIs têm no mínimo 1000 pb. A justificativa é que estaremos principalmente interessados no conteúdo gênico das RGIs; e é um fato empírico de genomas de procariotos que o tamanho médio de um gene é justamente 1000 pb.

2.11 Ilhas genômicas

De maneira geral, ilhas genômicas são regiões entre 10 e 100 mil pb que se diferenciam por diversas características do resto do genoma onde se encontram. É comum que um determinado genoma de uma espécie tenha uma ilha, ao passo que outro genoma da mesma espécie não a tenha. Em geral a origem dessas regiões é o processo de transferência horizontal, ou seja, essas regiões são adquiridas de outros organismos. Esse tipo de processo tem grande importância biológica, já que as as ilhas genômicas adquiridas por um organismo podem trazer vantagens evolutivas [20].

As ilhas genômicas tem uma relação forte com as RGIs, já que ilhas genômicas podem ser identificadas como RGIs também, caso existam no MAG e não existam no genoma de referência (e vice-versa). No entanto, outras regiões que não são ilhas genômicas também podem ser identificadas como RGIs. RGI é um conceito mais amplo que ilhas genômicas, já que não têm as mesmas restrições de definição que uma ilha genômica. Por outro lado, convém observar que caso a ilha genômica exista tanto no MAG como no genoma de referência, ela não será identificada como RGI.

2.12 Anotação funcional de genomas

Uma vez obtido o genoma, é possível identificar quais são seus genes, e por comparação de similaridade com banco de dados de genes já analisados anteriormente, podemos inferir qual é sua função dentro do organismo. Esse processo de identificação da função de um gene é chamado de anotação funcional. Diversos softwares realizam a identificação e anotação de genes em um genoma, como o PROKKA [21] e o PGAP [22].

2.13 COG

Para facilitar o processo de comparação dos genes, é possível identificar os genes por grupos funcionais, já que muitas vezes genomas a serem comparados tiveram seus genes anotados por diferentes ferramentas, com diferentes nomenclaturas. Comparar genes utilizando um banco de dados padronizado permite identificar genes compartilhados entre todos os genomas, além de facilitar a identificação da função dos genes que um genoma contém e outro não.

O banco de famílias gênicas COG (*Clusters of Orthologous Genes*) fornece 4877 famílias de genes classificadas de acordo com sua função [23]. Ao realizar o processo de anotação de um genoma contra esse banco de famílias, é possível identificar quais grupos os genes do genoma pertencem. Um gene é classificado como pertencente a um COG se o mesmo for similar aos já existentes no banco de dados. Na página dos COGs (https://www.ncbi.nlm.nih.gov/research/cog) existem agrupamentos por categorias e por *pathways* (processos compostos que envolvem diversos genes, como processos metabólicos). Caso uma das categorias/*pathways* seja do interesse de uma pesquisa/análise, após identificar quais COGs existem em seu genoma, um caminho comum é realizar o cruzamento entre a lista de COGs da categoria de interesse e os identificados no genoma.

2.14 CAZy

Além de bancos gerais de famílias gênicas como o COG, existem outros mais específicos para algumas funções. Um deles é o CAZy (*Carbohydrate-Active enZYmes Database*) [24], que contém um banco de dados de família de genes específicos para degradação, modificação e criação de enzimas de carboidratos. Uma das aplicações desse tipo de banco de dados é quando se deseja analisar processos de degradação de matéria orgânica. Como uma das motivações para realização deste trabalho foi a pesquisa de microbiomas da compostagem [25], que é um ambiente onde ocorre um intenso processo de degradação de matéria orgânica, a ferramenta proposta nesta dissertação também utiliza esse tipo de anotação.

2.15 ANI

ANI (*average nucleotide identity*) é uma medida de comparação de genomas simples e largamente utilizada. Como o nome diz, a medida ANI nos dá o número de nucleotídeos idênticos entre dois genomas, portanto supondo um alinhamento entre eles, dividindo esse número pelo tamanho do alinhamento. ANI é expressa na forma de porcentagem. Para se considerar dois genomas da mesma espécie é esperado um resultado de pelo menos 95% de ANI [26].

2.16 Pangenoma

Pangenoma é um conjunto contendo todos os genes de um grupo de genomas, geralmente da mesma espécie, e é mais um recurso muito utilizado para comparar genomas [27]. Utilizamos o pangenoma com o objetivo de identificar, dado um grupo de genomas:

- Genes principais (*core*): Genes compartilhados entre todos os genomas do grupo;
- Genes acessórios/compartilhados (*acessory/shared*): Genes compartilhados entre pelo menos dois genomas, mas não compartilhados entre todos os genomas;
- Genes específicos (*specific*): Genes que existem apenas em um genoma do grupo.

Na Figura 2.2 é possível visualizar uma das formas de representar o pangenoma, utilizando um diagrama de Venn.

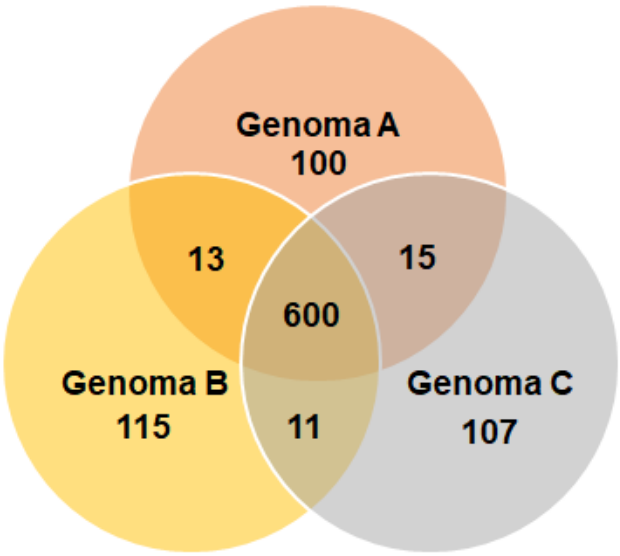


Figura 2.2: *Exemplo de representação do pangenoma*

Neste exemplo são comparados 3 genomas. Cada círculo representa um genoma. A intersecção entre todos os círculos representa os genes compartilhados entre todos os genomas (600 genes). As intersecções entre dois círculos representam os genes compartilhados somente entre dois genomas, e os genes específicos de cada genoma são representados na área do círculo sem intersecção.

Capítulo 3

Trabalhos relacionados

3.1 Plataformas para análise comparativa de genomas

VEuPathDB [28] (*Eukaryotic Pathogen, Vector and Host Informatics Resource*) é uma plataforma online que permite, entre suas diversas funções, realizar buscas avançadas utilizando diversos parâmetros de genes, genomas e anotações. A ferramenta tem um banco de dados já integrado, e seu foco é analisar e comparar genomas patogênicos eucarióticos. Já o MAGset tem como objetivo analisar genomas de bactérias. O trabalho é citado aqui pois sua ferramenta avançada de buscas compostas foi a inspiração para a busca de genes disponibilizada dentro do software MAGset. Convém salientar que a implementação dentro do MAGset é muito mais simples que a implementada pelo VEuPathDB. Na Figura 3.1 é possível ver um exemplo de busca composta realizada na ferramenta.

Figura 3.1: *Exemplo de busca composta na plataforma VEuPathDB*
Cada retângulo representa uma busca realizada no banco de genes. Ao realizar duas buscas diferentes, como mostrado acima, é possível mesclar os resultados. No exemplo, o usuário realizou duas buscas por genes: a primeira retornou 2.814 genes, a segunda 58.486 genes, mas apenas os resultados encontrados nas duas buscas (663 genes) serão retornados (intersecção dos resultados).

IMG/M [29] (*Integrated Microbial Genomes & Microbiomes*) é uma plataforma online que contém diversas funcionalidades para anotações, análises e comparações entre genomas e genes. Milhares de genomas já estão disponíveis na plataforma para análise, e é possível enviar os próprios genomas para realizar as comparações, no entanto, após um período definido na plataforma, o genoma que o usuário enviar deverá se tornar público.

Os genomas disponibilizados na base do IMG/M já estão anotados, sendo possível realizar buscas de genes por anotação e outras características em toda a base disponível.

PATRIC [30] (*Pathosystems Resource Integration Center*) é uma plataforma semelhante a IMG, permitindo diversos tipos de análises de genomas de procariotos e de vírus. Seu foco

principal é em patógenos, mas suporta análises de micro-organismos no geral. Assim como no IMG/M, a plataforma contém milhares de genomas pré-analisados e anotados. A ferramenta oferece comparações detalhadas ao nível de gene, inclusive permite a comparação de um gene em especifico entre genomas, exibindo a região adjacente ao gene para complementar a comparação. Ao enviar um genoma, é possível utilizar a função de busca por genoma similar (opção "*Similar genome finder*"), e assim identificar possíveis genomas da mesma espécie disponíveis na plataforma, e a partir daí realizar as comparações desejadas. Além das comparações entre genomas, também é disponibilizado como serviço online a montagem de genomas a partir dos *reads* das amostras.

MiGA [31] (*Microbial Genomes Atlas Online*) é uma plataforma online, mas também disponibiliza uma versão para instalação local. O software aceita como entrada tanto o genoma já montado, para então realizar as comparações, como o envio de amostras para realizar dentro da ferramenta todo o processo de montagem de genomas. Uma vez que o genoma está dentro da plataforma, é possível fazer a identificação de genes, avaliação da qualidade do genoma, anotações funcionais, identificação da espécie do genoma e identificações de quais genomas já existentes na base da plataforma são similares ao genoma do usuário. Caso existam genomas similares, também é possível gerar o pangenoma pela ferramenta.

As ferramentas já existentes são robustas e muito completas na parte comparação e anotação de genomas, mas elas não tem como foco encontrar regiões genômicas que existem no MAG mas não existem em nenhum genoma de referência (definidas neste trabalho como RGIs positivas) e regiões genômicas que existem em pelo menos um genoma de referência mas não existem no MAG (definidas neste trabalho como RGIs negativas).

O software proposto neste trabalho não tem como objetivo competir com as plataformas descritas, mas sim complementá-las ao realizar comparações para destacar as especifidades do MAG, e se possível utilizar essas diferenças para melhorar a qualidade da sua montagem.

3.2 Artigos sobre limitações de MAGs

A importância de MAGs mais completos, suas limitações e métodos para alcançar melhores genomas são discutidos em [32]. O artigo também traz informações relevantes sobre o processo de agrupamento de *contigs/scalffolds* (*binning*), descrevendo como esse processo pode causar erros no MAG, tanto inserindo sequências que não fazem parte do MAG como perdendo sequências que deveriam fazer parte do MAG.

O artigo também descreve diversos passos e propostas para melhorar a qualidade de montagem de MAGs, com o objetivo de completar a montagem de um MAG, entre eles:

- Curadoria da montagem usando GapFiller [33] para tentar aumentar/mesclar os *scaffolds*;
- Análise de regiões erroneamente repetidas;
- Análise do padrão do conteúdo GC;
- Checagem manual de cobertura;
- Verificação da existência de genes únicos universais (*single core genes - SCG)*;

Como descrito no próprio artigo, essa é uma tarefa árdua e demorada, e não existem softwares que executem estes passos de forma integrada.

Problemas similares na montagem de MAGs foram encontrados em [34]. Os autores indicam como problemas comuns na montagem de MAGs:

- Regiões repetidas que não são corretamente representadas;
- Falta de RNAs ribossomais;
- Falta de RNAs de transferência;
- Falta de elementos móveis;

Para se chegar nessas conclusões, os autores compararam MAGs obtidos de uma cultura de laboratório com genomas montados a partir isolados derivados dessa mesma cultura, além de comparar MAGs de outro estudo (amostras de oceano) com referências da mesma espécie publicados anteriormente a partir de isolados. Para encontrar as diferenças entre os MAGs e os genomas de referência, os autores criaram um conceito para avaliar as regiões que não existiam nos MAGs mas que existiam nas referências, chamado NRs (*not-binned regions* - regiões não inseridas no MAG), conceito muito similar ao definido neste trabalho: RGIs negativas.

Em outro artigo seguindo a mesma linha de avaliar problemas na montagem de MAGs [35], os autores compararam montagens obtidas a partir de uma amostra metagenômica de fezes com montagens de genomas de isolados, derivados da mesma amostra inicial. Com essa comparação, foi possível identificar que mesmo MAGs identificados com completude alta (em torno de 95%) por softwares de checagem como o CheckM [19] continham apenas 77% do genes considerados *core* e 50% dos genes acessórios, em média. O artigo também aponta que a contaminação é subestimada, já que enquanto o reportado pelo CheckM foi de 1,5% de contaminação, foi identificado que 5% dos genes constantes nos MAGs não estavam nos isolados.

Em nenhum dos artigos aqui citados é fornecida uma ferramenta para identificar regiões (e quais genes contidos em cada região) que existem no MAG e não existem no genoma de referência (regiões que merecem atenção pois podem ser contaminações) e as regiões que existem no genoma de referência e não existem no MAG (regiões faltantes). Caso alguém queira identificar essas regiões em seus genomas, terá que fazer de forma manual.

Outro importante tópico não abordado nos dois artigos aqui citados é: As regiões faltantes nos MAGs mas existentes no genoma de referência estão nos *reads* das amostras metagenômicas? Se sim, seria possível utilizá-los para melhorar a completude do genoma? Encontrar as regiões faltantes nas amostras é uma das funcionalidades da ferramenta proposta neste trabalho, como se verá no próximo capítulo. A utilização destas regiões para a melhorar o MAG também foi abordada neste trabalho (Capítulo 5), com os resultados obtidos descritos no Capítulo 7.

Capítulo 4

MAGset

Neste capítulo apresentamos nossa principal contribuição, o software MAGset, um comparador de MAGs com genomas de referência da mesma espécie. O objetivo do MAGset é fornecer ao usuário as especificidades do MAG ao compará-lo com outros genomas. Ainda que o objetivo principal seja comparar MAGs com genomas de referência que sejam completos, também é possível utilizá-lo comparando com outros genomas incompletos, ou mesmo com outros MAGs.

Após a execução do software, os seguintes resultados são apresentados:

- Matriz ANI comparando todos os genomas;
- Pangenoma;
- Anotações dos genes com o banco de dados CAZy e COG;
- Regiões genômicas de interesse;
- Lista de genes para análise dos dados via buscas avançadas dos genes por suas características;
- Resultado da validação das RGIs negativas contra as amostras que originaram o MAG (MAGcheck);

O usuário deve inserir os genomas, tanto o MAG como as referências, como entrada no software. O software suporta a comparação de um MAG com um ou mais genomas de referência da mesma espécie. Algumas funções dependem do fornecimento dos genomas já anotados para serem executadas (pangenoma, anotações com CAZy e COG e disponibilização da busca avançada de genes). A execução do módulo MAGcheck depende do fornecimento dos *reads* das amostras de onde o MAG foi originalmente obtido.

O *pipeline* de execução do software é exibido na Figura 4.1, e um exemplo dos resultados gerados após a análise é exibido na Figura 4.2.

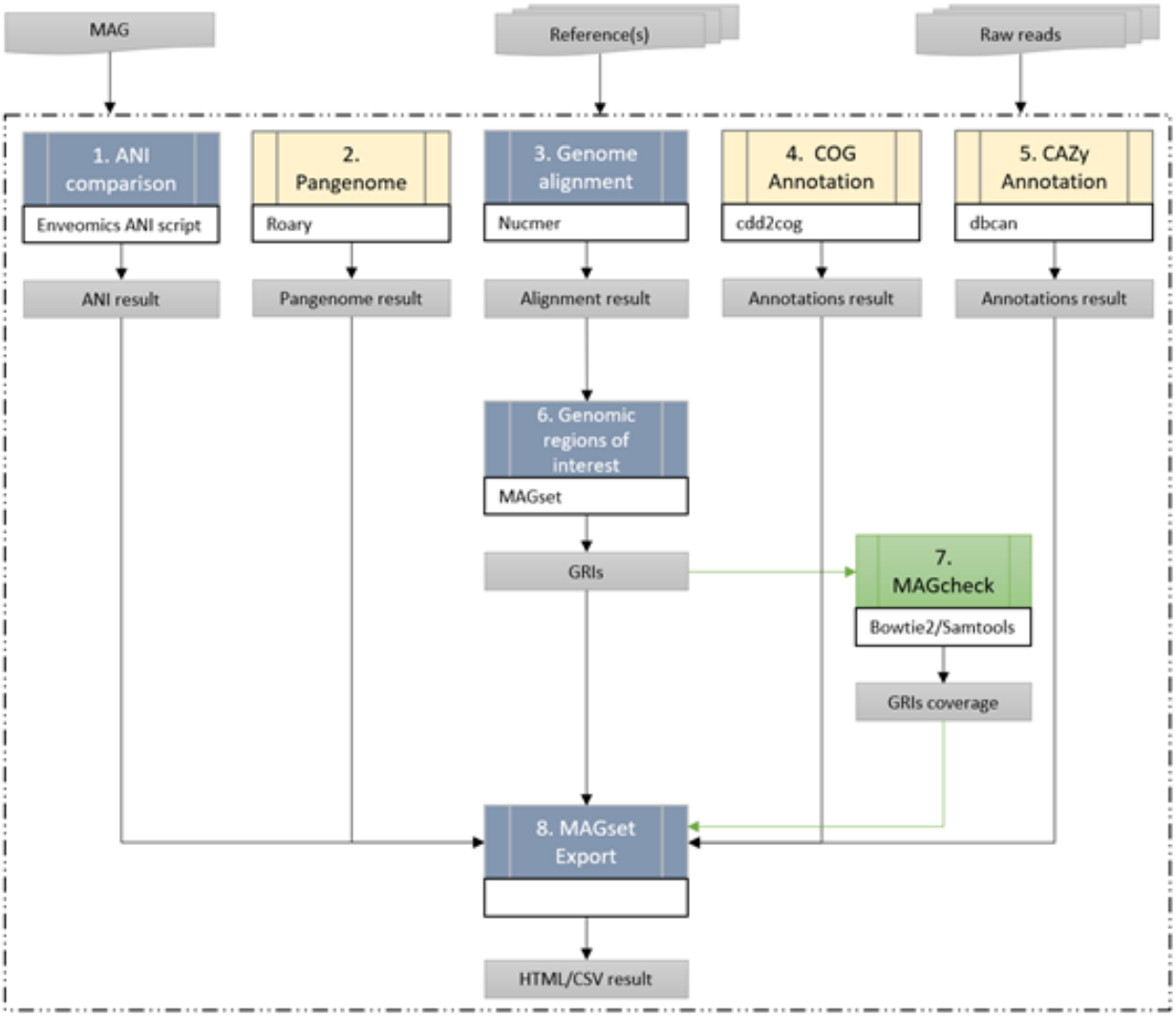
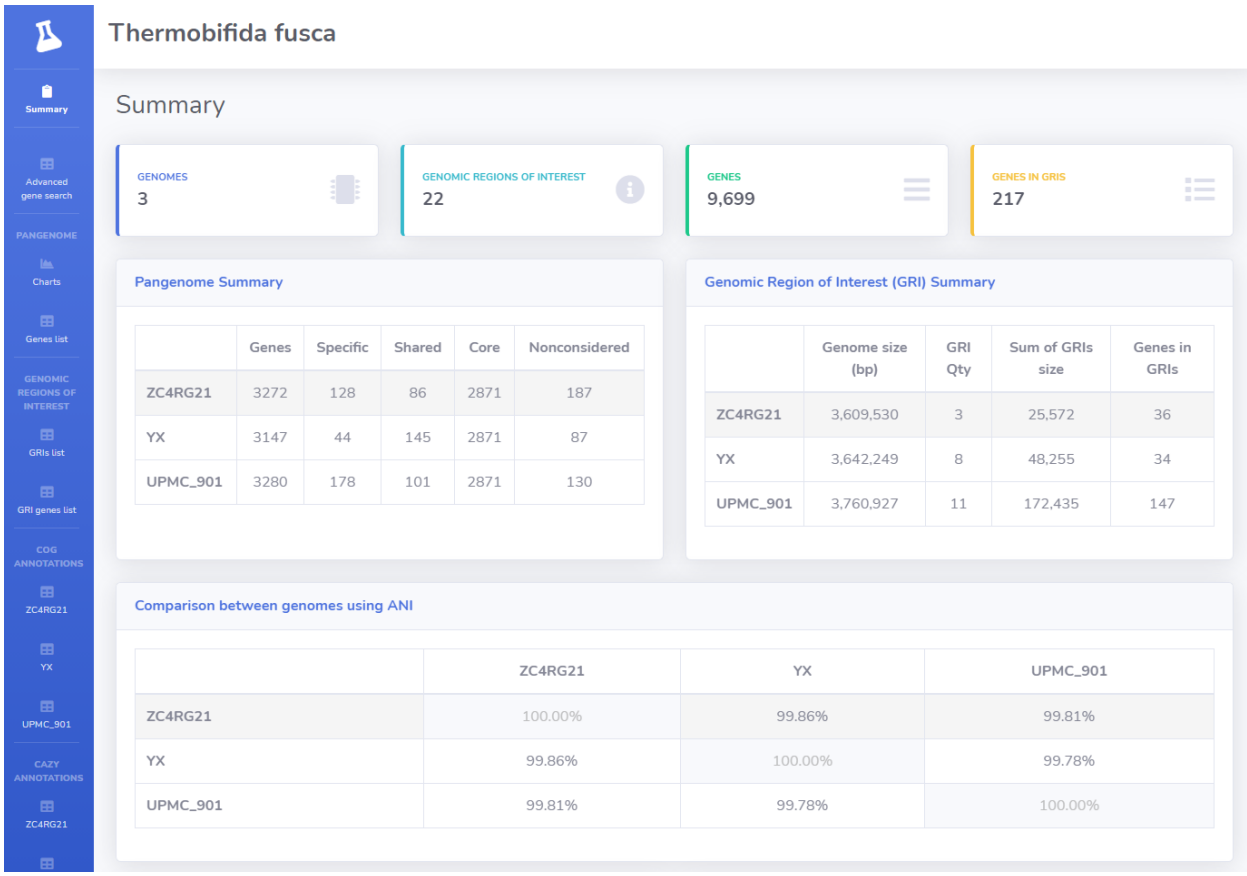


Figura 4.1: *Pipeline de execução do MAGset*
Os itens em azul são sempre executados, enquanto os itens em amarelo só são executados quando a entrada é um arquivo no formato GenBank (.gbff). O item "7. MAGcheck", em verde, só é executado quando os *reads* da amostra de onde o MAG foi extraído são fornecidos.



	Genes	Specific	Shared	Core	Nonconsidered
ZC4RG21	3272	128	86	2871	187
YX	3147	44	145	2871	87
UPMC_901	3280	178	101	2871	130

	Genome size (bp)	GRI Qty	Sum of GRIs size	Genes in GRIs
ZC4RG21	3,609,530	3	25,572	36
YX	3,642,249	8	48,255	34
UPMC_901	3,760,927	11	172,435	147

	ZC4RG21	YX	UPMC_901
ZC4RG21	100.00%	99.86%	99.81%
YX	99.86%	100.00%	99.78%
UPMC_901	99.81%	99.78%	100.00%

Figura 4.2: *Exemplo da página inicial dos resultados do MAGset*

4.1 Metodologia

4.1.1 Identificação de regiões genômicas de interesse (RGI)

Para identificar as RGIs, o MAGset utiliza a ferramenta nucmer [36], com o parâmetro *maxmatch*. São feitas duas comparações:

- MAG versus os genomas de todas as referências informadas: Em busca das regiões exclusivas do MAG, o resultado do nucmer é utilizado para encontrar quais regiões do MAG não existem em nenhuma das referências. Como o resultado do nucmer retorna quais são as áreas similares encontradas, todas as regiões em que não houve "*match*" com a referência são consideradas RGIs positivas, desde que tenham o tamanho mínimo de 5.000 pb.

- Referência versus MAG: Em busca de regiões não existentes no MAG mas existente em uma ou mais referências, o processo é similar ao descrito anteriormente, com a diferença que a execução do nucmer é realizada para cada referência versus o MAG. Todas as regiões em que não houve "*match*" são consideradas RGIs negativas, desde que tenham o tamanho mínimo de 5.000 pb.

O tamanho mínimo de 5.000 pb foi escolhido por ser suficientemente grande para conter regiões com possíveis genes de interesse na análise, mas não tão pequeno a ponto de trazer um excesso de pequenas regiões diferentes entre os genomas sem muito valor biológico. Este tamanho foi definido para as comparações realizadas neste trabalho e é o valor padrão de

execução do software. Esse valor pode ser modificado pelo usuário, caso o mesmo entenda que a RGI deva ser maior ou menor na sua análise.

Os resultados do nucmer passam pelo seguinte filtro para serem considerados: Identidade mínima de 90% e tamanho mínimo de 1.000 pb. Esses valores foram escolhidos para evitar que um pequeno alinhamento desconsiderasse uma RGI. Também são valores que podem ser configurados de forma diferente pelo usuário do sistema, caso considere necessário.

Após a identificação das RGIs negativas, estas são agrupadas se, ao compará-las entre si com o nucmer, tiverem similaridade em pelo menos 80% das bases (valor padrão, configurável na execução do software). Esse agrupamento é útil para identificar RGIs negativas de diferentes genomas de referência que são iguais ou muito similares.

Na Figura 4.3 e 4.4 é possível visualizar como o software irá exibir o resultado após a análise. Além do resultado em HTML exibido nas figuras, o resultado também é disponibilizado em CSV.

Genomic Region of Interest (GRI) Summary

	GRI Qty	Genes in RGIs
MAG	2	N/A
REF_A	2	N/A
REF_B	2	N/A

Figura 4.3: *Resultado gerado pelo MAGset identificando as RGIs (resumo)*

GRIs

Show 10 entries Search:

Id	Genome	Parent	Size	Start position	End position
NGRI0001_01	REF_A	sample_ref01_contig_01	10,001	6,001	16,001
NGRI0002_01	REF_A	sample_ref01_contig_01	7,500	18,501	26,000
NGRI0002_02	REF_B	sample_ref02_contig_05	7,500	1	7,500
NGRI0003_01	REF_B	sample_ref02_contig_01	10,000	1	10,000
PGRI0004_01	MAG	sample_mag_contig_02	6,000	1	6,000
PGRI0005_01	MAG	sample_mag_contig_03	7,500	1	7,500
Id	Genome	Parent	Size	Start position	End position

Showing 1 to 6 of 6 entries Previous 1 Next

Export results to CSV

Figura 4.4: *Resultado gerado pelo MAGset identificando as RGIs (detalhado)*
Para cada RGI são exibidos o tamanho, a posição inicial e a posição final. Caso a RGI seja a mesma em dois genomas diferentes, elas têm o mesmo código, como é o caso da RGI 0002 nesta imagem.

4.1.2 Matriz de similaridade com as referências, utilizando ANI

Para permitir uma interpretação mais adequada dos resultados obtidos, MAGset também gera uma matriz com a similaridade dos genomas comparados, utilizando o script ani.rb [37] para realizar as comparações. A Figura 4.5 permite visualizar um exemplo do resultado apresentado.

Comparison between genomes using ANI

	MAG	REF_A	REF_B
MAG	100.00%	98.57%	99.30%
REF_A	98.57%	100.00%	98.93%
REF_B	99.30%	98.93%	100.00%

Figura 4.5: *Resultado gerado pelo MAGset comparando genomas com ANI*

4.1.3 Comparação dos genomas via pangenoma

Caso os genomas enviados para comparação estejam no formato GenBank (.gbff), também será gerado o pangenoma, utilizando a ferramenta Roary [38]. Como é necessário converter o arquivo .gbff para .gff (entrada da ferramenta Roary), o script bp_genbank2gff3.pl [39] é utilizado para esta tarefa. Os resultados do pangenoma são exibidos de 3 formas:

- Página inicial: resumo dos resultados do pangenoma, com quantidade de genes específicos/"*specific*" (apenas esse genoma contém esses genes), compartilhados/ *"shared"* (mais de um genoma contém esses genes, mas não todos), principais/"*core*" (genes compartilhados entre todos os genomas) e genes desconsiderados/"*nonconsidered*" (genes incompletos e outros genes desconsiderados pelo Roary) (Figura 4.6).
- Gráficos: Os gráficos são gerados com o script roary_plots.py, disponibilizados pelo programa Roary (Figura 4.7);
- Lista detalhada dos genes identificados pelo pangenoma, indicando de quais genomas ele faz parte (Figura 4.8);

Pangenome Summary

	Genes	Specific	Shared	Core	Nonconsidered
ZC4RG21	3272	128	86	2871	187
YX	3147	44	145	2871	87
UPMC_901	3280	178	101	2871	130

Figura 4.6: *Resultado gerado pelo MAGset: Pangenoma (resumo)*
Resultado do pangenoma exibido na página inicial do MAGset. O MAG sempre é exibido com o fundo em cinza, para facilitar a identificação. Nesse exemplo é utilizado o MAG do projeto Metazoo, da espécie *Thermobifida fusca*, identificado como ZC4RG21, comparado com duas referências completas: cepa YX (GCF_000012405.1) e cepa UPMC_901 (GCF_015034585.1).

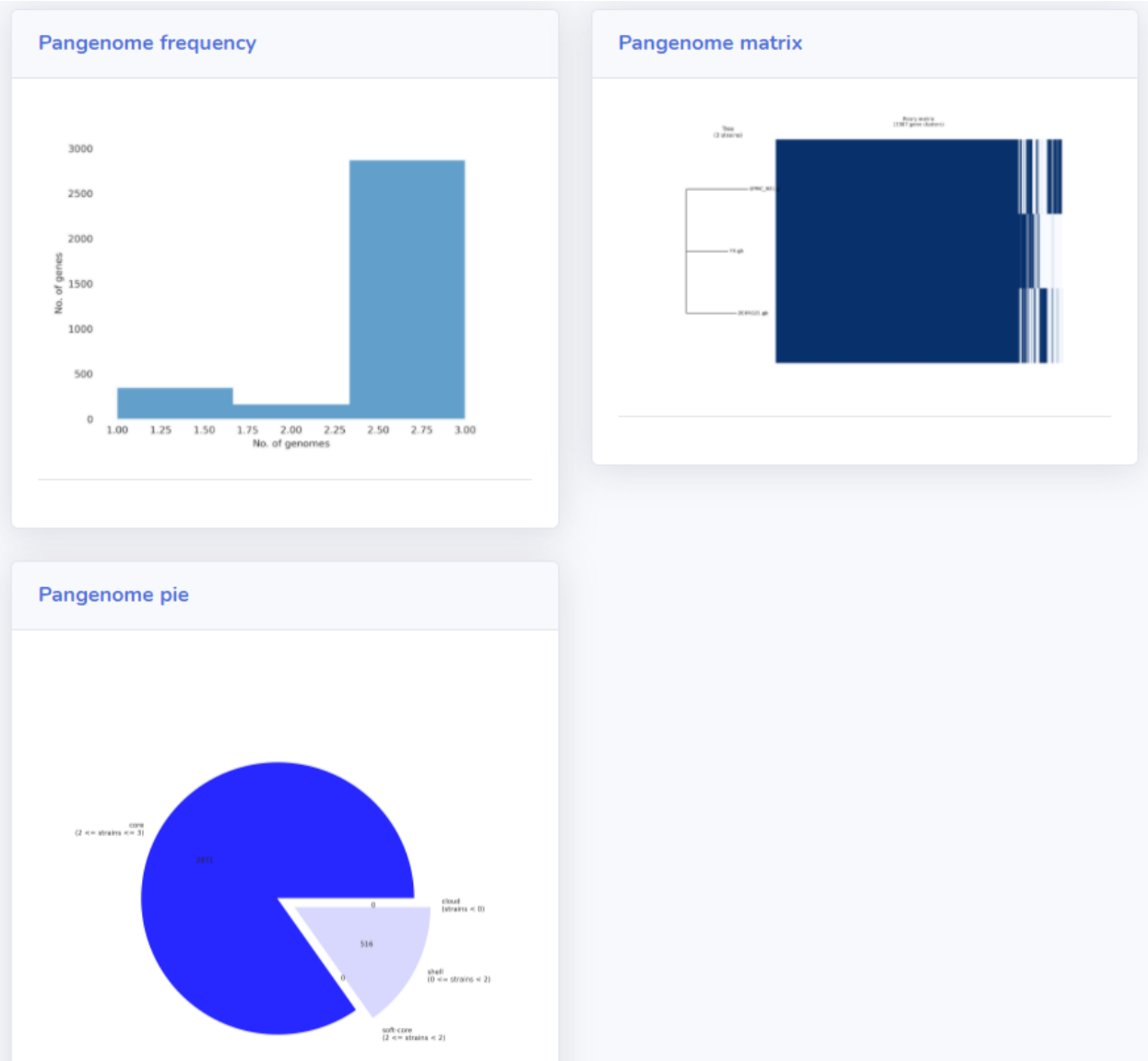


Figura 4.7: *Resultado gerado pelo MAGset: Pangenoma (gráficos)*
Gráficos gerados do pangenoma utilizando o Roary e o script roary_plots.py.

Genes

Show 10 entries Search:

Gene name	Annotation	N. isolates	UPMC_901	YX	ZC4RG21
aceA	isocitrate lyase	3	✓	✓	✓
acnA	aconitate hydratase AcnA	3	✓	✓	✓
acs	acetate--CoA ligase	3	✓	✓	✓
alaS	alanine--tRNA ligase	3	✓	✓	✓
ald	alanine dehydrogenase	3	✓	✓	✓
amt	ammonium transporter	3	✓	✓	✓
arc	proteasome ATPase	3	✓	✓	✓
argB	acetylglutamate kinase	3	✓	✓	✓
argF	ornithine carbamoyltransferase	3	✓	✓	✓
argH	argininosuccinate lyase	3	✓	✓	✓
Gene name	Annotation	N. isolates	UPMC_901	YX	ZC4RG21

Showing 1 to 10 of 3,387 entries Previous 1 2 3 4 5 ... 339 Next

Export results to CSV

Figura 4.8: *Resultado gerado pelo MAGset: Genes do pangenoma*
Lista de genes analisados no pangenoma, indicando em qual genoma cada gene aparece, obtidos a partir dos resultados do software Roary.

4.1.4 Anotações COG e CAZy

Caso os genomas enviados para comparação estejam no formato GenBank (.gbff), eles serão anotados utilizando o banco de dados COG [23] e CAZy [24]. A anotação utilizando o banco de dados CAZy pode ser desabilitada pelo usuário, caso ela não seja necessária. As anotações são apresentadas em duas visualizações nos resultados gerados:

- Página inicial: resumo das anotações obtidas por cada genoma e por cada tipo de anotação (Figura 4.9)).
- Lista detalhada dos genes anotados (Figura 4.10)), com uma página de resultado por genoma e por tipo de anotação (COG e CAZy);

Annotations Summary

	CAZy	COGs
ZC4RG21	93	1778
YX	96	1788
UPMC_901	93	1812

Figura 4.9: *Resultado gerado pelo MAGset: genes anotados com CAZy e COG (resumo)*

COGs

Show 10 entries Search:

Locus tag	COG ID	COG Description
DIU53_00005	COG1703	Putative periplasmic protein kinase ArgK and related GTPases of G3E family
DIU53_00010	COG1884	Methylmalonyl-CoA mutase, N-terminal domain/subunit
DIU53_00015	COG1884	Methylmalonyl-CoA mutase, N-terminal domain/subunit
DIU53_00025	COG0386	Glutathione peroxidase
DIU53_00035	COG1739	Uncharacterized conserved protein
DIU53_00055	COG0860	N-acetylmuramoyl-L-alanine amidase
DIU53_00080	COG0150	Phosphoribosylaminoimidazole (AIR) synthetase
DIU53_00085	COG0034	Glutamine phosphoribosylpyrophosphate amidotransferase
DIU53_00095	COG1392	Phosphate transport regulator (distant homolog of PhoU)
DIU53_00100	COG0306	Phosphate/sulphate permeases
Locus tag	COG ID	COG Description

Showing 1 to 10 of 1,778 entries Previous 1 2 3 4 5 ... 178 Next

Export results to CSV

Figura 4.10: *Resultado gerado pelo MAGset: genes anotados com COG (detalhado)* Lista de genes anotados do genoma ZC4RG01 utilizando o banco de dados COG. Para cada genoma, é gerada uma página de resultados como essa. O mesmo ocorre com as anotações CAZy.

4.1.5 Busca avançada de genes

Caso os genomas enviados para comparação estejam no formato GenBank (.gbff), é possível utilizar a busca avançada de genes pelo MAGset. A busca avançada é inspirada na funcionalidade de buscas cumulativas da plataforma VeuPathDB [28], que permite realizar diferentes buscas para adicionar, interseccionar ou remover resultados entre elas. Exemplo de buscas possíveis na ferramenta:

- Genes anotados com o banco de dados COG **e** genes que são específicos **e** genes do MAG **e** genes que não estão em RGIs;
- Genes anotados com o banco de dados CAZy **e** que não estão em um dos genomas;

Na Figura 4.11 é possível visualizar um exemplo de busca de genes utilizando a ferramenta.

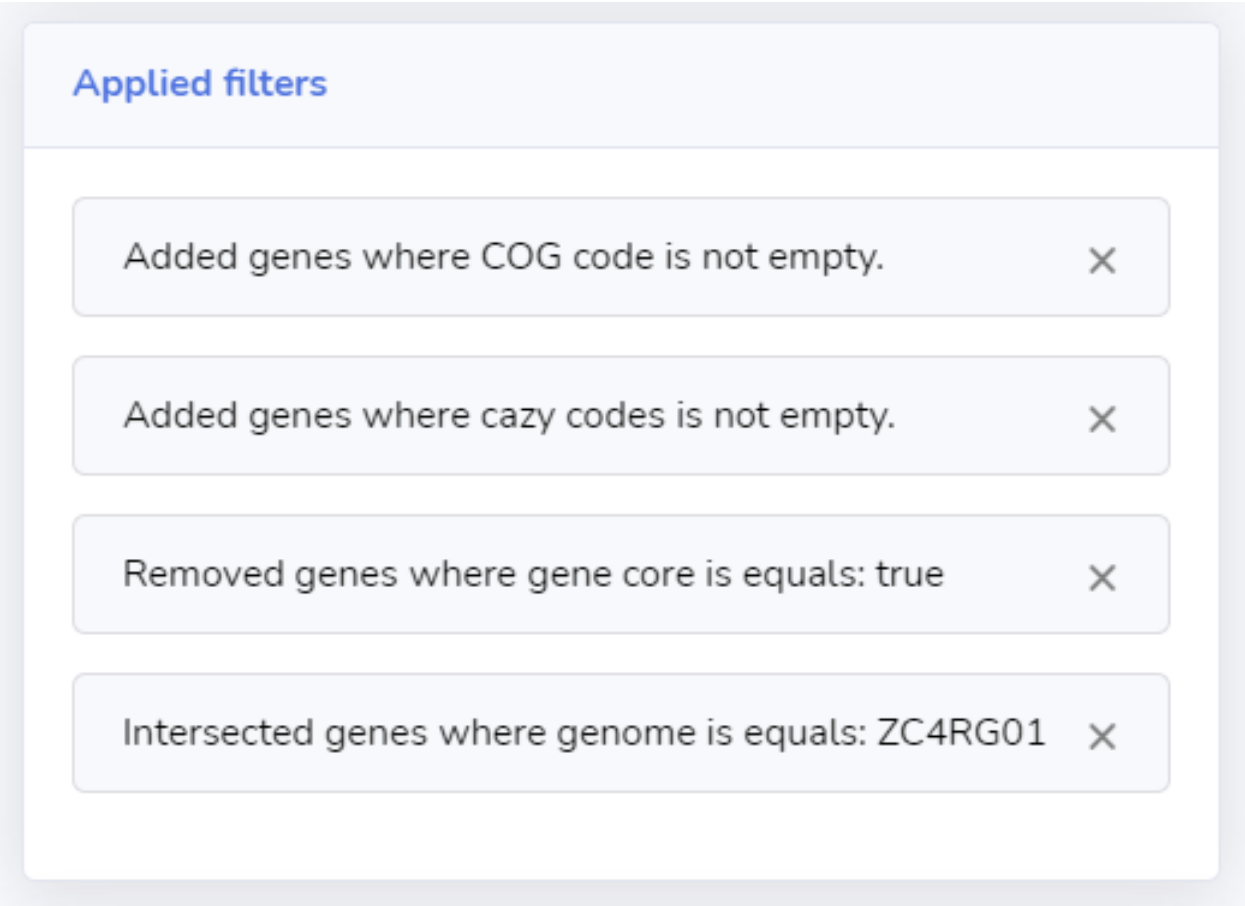


Figura 4.11: *Exemplo de busca de genes no MAGset: filtros aplicados*
Exemplo de filtros que podem ser aplicados utilizando a ferramenta. Foram aplicados 4 filtros: Adicione todos os genes com código COG diferente de vazio, adicione todos os genes com código CAZy diferente de vazio, do resultado obtido remova todos os genes *core* e faça uma intersecção do resultado com genes que o genoma seja igual a ZC4RG01 (nesse caso, o MAG em análise).

4.1.6 MAGcheck: Verificação de RGIs negativas

Além de comparar os genomas, é possível também verificar se as RGIs negativas (RGIs que existem em uma ou mais referências, mas não existem no MAG) identificadas pelo

MAGset existem nas amostras metagenômicas originais, que deram origem ao MAG. Como o processo de obtenção de MAGs não é perfeito, é possível que existam partes do MAG não corretamente identificadas, apesar de estarem disponíveis nas amostras.

Para identificar essas regiões perdidas nas amostras, foi criado o módulo MAGcheck. Para utilizá-lo, é necessário, além dos genomas que serão comparados, informar também os arquivos com *reads* das amostras utilizadas na montagem. O MAGcheck irá mapear os *reads* das amostras contra as RGIs negativas utilizando o software Bowtie2 [40] com o parâmetro "*very-sensitive*" para obter maior acurácia, e caso a cobertura de uma região seja maior que 80% (valor padrão, mas configurável pelo usuário), a RGI será destacada em amarelo nos resultados da ferramenta (Figura 4.13). Um resumo dos resultados encontrados é exibido na página inicial dos resultados (Figura 4.12).

MAGCheck Summary

	GRIs found in raw data	Sum of GRIs size	Genes in GRIs
REF_A	1	7,500	N/A
REF_B	1	7,500	N/A

The matched reads can be found at: ${output_folder}/07_magcheck/reads/

Figura 4.12: *Resultado gerado pelo MAGset: RGIs encontradas pelo MAGcheck (resumo)*

Os *reads* mapeados também são disponibilizados nos resultados gerados pelo MAGset, caso o usuário queira utilizá-los para complementar a montagem do MAG posteriormente, como descrito neste trabalho no Capítulo 5.

Uma premissa importante do módulo MAGcheck é que consideramos que os *reads* que alinham com as RGIs negativas pertencem ao genoma que originou o MAG, e não a outro genoma existente no mesmo ambiente de onde a amostra foi obtida. Consideramos que essa premissa é válida pois toda RGI negativa está presente em pelo menos um genoma de referência.

Cabe salientar que MAGs podem representar mosaicos de cepas da mesma espécie existentes na amostra se estas são muito parecidas [41] [42], e desta forma, as RGIs podem ter sido encontradas em cepas que não são as dominantes.

GRIs

Show 10 entries Search:

Id	Genome	Parent	Size	Start position	End position	Genes Qty	Covered positions	Covered positions (%)	Mean depth
NGRI0001_01	YX	NC_007333	5221	543892	549112	5	4313	82.61	17.47
NGRI0001_02	UPMC_901	NZ_CP063232	5726	1481050	1486775	7	5508	96.19	18.59
NGRI0002_01	YX	NC_007333	6200	1313191	1319390	3	6189	99.82	15.41
NGRI0002_02	UPMC_901	NZ_CP063232	6200	2250502	2256701	3	6189	99.82	15.41
NGRI0003_01	YX	NC_007333	7029	1529860	1536888	5	3906	55.57	40.41
NGRI0004_01	YX	NC_007333	6370	1753956	1760325	7	6314	99.12	22.26
NGRI0004_02	UPMC_901	NZ_CP063232	6369	2702201	2708569	7	6313	99.12	22.22
NGRI0005_01	YX	NC_007333	5317	2386899	2392215	3	5316	99.98	1147.05
NGRI0005_02	YX	NC_007333	7241	2779911	2787151	5	7241	100.00	851.85
NGRI0005_03	YX	NC_007333	5389	3039180	3044568	3	5352	99.31	1131.55
Id	Genome	Parent	Size	Start position	End position	Genes Qty	Covered positions	Covered positions (%)	Mean depth

Showing 1 to 10 of 21 entries Previous 1 2 3 Next

Figura 4.13: *Resultado gerado pelo MAGset: RGIs encontradas pelo MAGcheck (detalhado)* Exemplo de resultado com diversas RGIs negativas encontradas nas amostras utilizadas para a montagem do MAG. Todas as linhas em amarelo identificam RGIs negativas com pelo menos 80% de cobertura nos *reads* das amostras, apesar de não fazerem parte do MAG.

4.2 Arquitetura

O software MAGset foi construído para execução via linha de comando, e como depende de diversos outros softwares de bioinformática para ser executado, que por sua vez dependem de diversos outros softwares construídos em diferentes linguagens, optamos por distribuí-lo utilizando uma imagem em *container* Docker [43]. *Containers* facilitam o processo de instalação, evitando conflitos com outros softwares já instalados no ambiente do usuário e têm um bom nível de reprodutibilidade independentemente do ambiente que está sendo executado. A inspiração para a criação do software utilizando um *container* para sua distribuição foi baseada no PGAP [22], software distribuído pelo NCBI para anotação de genomas.

Utilizamos a linguagem *shell script* para realizar as chamadas aos diversos softwares do *pipeline* e a linguagem *Java* para construir o módulo de identificação das RGIs e exportação dos resultados.

Já para construir o resultado em HTML, utilizamos *Javascript*, CSS e HTML. Como base para o layout dos resultados, utilizamos um modelo fornecido gratuitamente em https://startbootstrap.com/theme/sb-admin-2.

4.3 Softwares utilizados no *pipeline*

Dentro do *pipeline*, além dos módulos desenvolvidos neste trabalho, são executados os seguintes softwares:

- Enveomics ANI [37] para a geração da matriz ANI entre os genomas;
- Roary [38] para gerar o pangenoma dos genomas;
- Fasttree [44] para a geração de arquivos auxiliares do pangenoma;
- Nucmer [36] para comparação das sequências;
- Bowtie2 [40] para alinhamento dos *reads* nos genomas;
- Dbcan2 [45] para anotação CAZy;
- cdd2cog [46] para anotação COG;
- Diversos softwares e bibliotecas auxiliares para leitura/conversão dos arquivos nos formatos necessários: BioJava [47] , BioPerl [39] , BioPython [48] , EMBOSS [49] e SAMtools [50].

4.4 Entradas suportadas

Os genomas a serem analisados devem ser de um dos dois formatos suportados: Arquivos FASTA ou arquivos GBFF (formato de arquivo para genomas anotados pelo NCBI). Todos os genomas analisados (MAG e referências) devem ser do mesmo tipo. A ferramenta foi criada para suportar a comparação de até 10 genomas. Embora o software não bloqueie comparações com mais genomas, esse é o valor máximo recomendado por questões de performance e tempo de execução, e foi a quantidade máxima que a ferramenta foi validada.

Os *reads* das amostras devem estar no formato FASTA ou FASTQ. *Reads* compactados também são suportados (na extensão .gz).

4.5 Formato dos resultados

Para permitir uma leitura dos resultados com boa usabilidade e facilitar a interpretação, os resultados são gerados em dois formatos: HTML e CSV (*comma separated values*). Enquanto a versão de resultados em CSV pode ser utilizada para servir como entrada para outros softwares, a versão em HTML permite uma navegação fluída e facilita a interpretação dos dados por apresentar os dados de forma sintética e analítica.

4.6 Disponibilidade

Futuramente, o software estará disponível para download no repositório GitHub, no endereço:

https://github.com/LaboratorioBioinformatica/MAGset.

Além da disponibilização do software pronto para o uso, também estará disponível seu código fonte e alguns tutoriais para facilitar o uso pelos novos usuários (também no GitHub):

- *Quick start*
- *Tutorial - GBFF files as input*
- *Tutorial - How to use the advanced gene search*
- *Tutorial - Using MAGcheck*

O software foi construído para ser executado em sistemas Linux, as versões validadas estarão descritas na documentação online dentro do GitHub, bem como o uso de memória e o tempo de execução esperados.

Capítulo 5

Metodologia do refinamento de MAGs

Apesar de não fazer parte do software, para validar os resultados e verificar se é possível melhorar os MAGs a partir dos resultados obtidos pelo MAGcheck, neste trabalho também realizamos a remontagem de MAGs. Todos os MAGs que tiveram pelo menos uma RGI negativa encontrada nos *reads* das amostras pelo módulo MAGcheck foram remontados. Para reconstruir os genomas, foram executados os seguintes passos:

- Obtenção dos *reads* resultantes do módulo MAGcheck, que alinharam com as RGIs negativas (disponibilizado pelo software MAGset);
- Obtenção dos *reads* que alinham com o MAG, utilizando Bowtie2 [40] parâmetro "*very-sensitive*";
- Remoção dos *reads* duplicados obtidos nos dois passos anteriores;
- Remontagem com os *reads* utilizando Spades [13] ou MegaHIT [14], caso os resultados com Spades não sejam bons (completude inferior ao MAG original ou contaminação acima de 10%);
  - Caso as amostras metagenômicas disponíveis já tenham sido corrigidas pelo Spades anteriormente, a remontagem foi feita com o parâmetro "*only-assembler*";
- Validação do MAG resultante com CheckM [19], observando a completude e contaminação com tamanho de *contigs* mínimos entre 1.000 e 2.500 pb, selecionando o melhor resultado.

Os resultados obtidos com esta metodologia estão descritos no Capítulo 7. O refinamento descrito acima não trouxe bons resultados (houve aumento da contaminação de forma exagerada) para a espécie *Desulfurellaceae bacterium* do projeto SRA. Nesse caso, foi utilizada a seguinte abordagem, com bons resultados (também descritos no Capítulo 7):

- Obtenção das sequências das RGIs negativas encontradas pelo MAGcheck;
- Obtenção de todos os *contigs* gerados no processo de montagem metagenômico (arquivo gerado no processo de montagem antes de executar o processo de "*binning*");
- Identificação dos contigs que contém um alinhamento de pelo menos 1.000 pb e uma identidade de pelo menos 90%, ao compará-lo com as sequências das RGIs negativas utilizando o programa nucmer;
- Mesclagem dos *contigs* do MAG original com os novos contigs encontrados;

- Validação do novo MAG resultante com CheckM, observando a completude e contaminação com tamanho de *contigs* mínimos entre 1.000 e 2.500 pb, selecionando o melhor resultado.

A segunda abordagem só pode ser utilizada nos artigos ou projetos em que todos os *contigs* gerados no processo de montagem estão disponíveis. Esses *contigs* não são normalmente disponibilizados nos dados publicados em artigos, portanto é uma abordagem mais comum quando o processo de melhoria da montagem for executada pelo próprio responsável pela criação do MAG. Esta abordagem é bem mais simples de ser executada comparada com a primeira abordagem aqui descrita, e pode trazer ganhos consideráveis no MAG em termos de tamanho e completude.

Capítulo 6

Conjunto de dados analisados

Para validar o funcionamento do software, foram utilizados MAGs de 5 artigos publicados e MAGs montados por outros estudantes do nosso laboratório:

- Dados do artigo [1], neste trabalho identificado apenas como "artigo FT" (sigla para *Fecal Transfer*) para facilitar a referência.
- Dados do artigo [2], neste trabalho identificado apenas como "artigo GI" (sigla para *Gut Infant*) para facilitar a referência.
- Dados do artigo [3], neste trabalho identificado apenas como "artigo HS" (sigla para *Hot Springs*) para facilitar a referência.
- Dados do artigo [4], neste trabalho identificado apenas como "artigo PP" (sigla para *Phytoplankton*) para facilitar a referência.
- Dados do artigo [5], neste trabalho identificado apenas como "artigo SG" (sigla para *Shifts from Human Gut*) para facilitar a referência.
- Dados do projeto Metazoo: fezes de bugios, neste trabalho identificado apenas como "Bugios" para facilitar a referência.
- Dados do projeto Metazoo: Composto maduro, neste trabalho identificado apenas como "Composto Maduro" para facilitar a referência.
- Dados do projeto Metazoo: Inóculo, neste trabalho identificado apenas como "Inóculo" para facilitar a referência.
- Dados do projeto Metazoo: Lago, neste trabalho identificado apenas como "Lago" para facilitar a referência.
- Dados do projeto Metazoo: Composteira ZC3, neste trabalho identificado apenas como "ZC3" para facilitar a referência.
- Dados do projeto Metazoo: Composteira ZC4, neste trabalho identificado apenas como "ZC4" para facilitar a referência.
- Dados do projeto Sistema Recifal Amazônico, sobre o microbioma de esponjas, neste trabalho identificado apenas como "SRA" para facilitar a referência.

Todos esses trabalhos contêm MAGs de espécies conhecidas e os dados metagenômicos para validá-los com o MAGcheck e remontá-los. Os detalhes das amostras metagenômicas e as ferramentas utilizadas para obter os MAGs nos projetos e artigos podem ser consultados no Apêndice A.

A Tabela 6.1 contém a lista de MAGs que foram analisados nesse trabalho e com quais referências cada MAG foi comparado. As referências foram obtidas no NCBI. Apesar de preferencialmente a comparação de MAGs ser realizada contra genomas de isolados, abrimos uma exceção para os MAGs do projeto SRA, já que os MAGs obtidos neste projeto não têm genoma de referência a partir de isolados publicados. Encontramos MAGs muito similares publicados, e como exemplo das possibilidades de comparação utilizando o MAGset, inserimos esses dados também.

Tabela 6.1: *Lista de MAGs analisados*

Origem	Espécie	Completude (%)	Contaminação (%)	Genoma(s) de referência
Artigo FT	*Bacteroides uniformis*	92,55	1,73	GCF_006742345.1
Artigo FT	*Bacteroides vulgatus*	79,53	8,56	GCF_000012825.1
Artigo GI	*Clostridium baratii*	99,19	1,61	GCF_001991075.2 GCF_000789395.1
Artigo GI	*Enterococcus faecalis*	99,63	0	GCF_000391485.2
Artigo GI	*Enterococcus faecium*	98,19	0,47	GCF_001720945.1
Artigo HS	*Ardenticatena maritima*	70,22	3,31	GCF_001306175.1
Artigo HS	*Chloroflexus aurantiacus*	92,14	0,94	GCF_000022185.1
Artigo HS	*Thermosynechococcus sp*	97,96	0,94	GCF_000505665.1
Artigo PP	*Marivivens sp*	95,36	0,95	GCF_001908835.1
Artigo PP	*Ruegeria pomeroyi*	97,56	2,18	GCF_000011965.2
Artigo SG	*Enterococcus faecalis*	99,63	0,19	GCF_000391485.2
Artigo SG	*Enterococcus faecium*	99,63	0,94	GCF_001720945.1
Bugios	*Prevotella sp*	99,26	1,11	GCF_002251295.1
Bugios	*Treponema berlinense*	87,57	0,93	GCF_900167025.1
Bugios	*Treponema succinifaciens*	90,05	1,93	GCF_000195275.1
Composto Maduro	*Mycolicibacterium thermoresistibile*	80,59	3,07	GCF_900187065.1
Composto Maduro	*Novibacillus thermophilus*	60,66	2,69	GCF_002005165.1
Composto Maduro	*Planifilum fulgidum*	62,72	5,67	GCF_900113175.1
Composto Maduro	*Thermocrispum municipale*	80,59	3,07	GCF_000427825.1
Inóculo	*Sphaerobacter thermophilus*	68,80	4,52	GCF_000024985.1
Lago	*Limnohabitans sp*	97,27	3,8	GCF_001412575.1
ZC3	*Planifilum fulgidum*	93,08	0	GCF_900113175.1
ZC3	*Thermobifida fusca*	99,45	0	GCF_015034585.1 GCF_000012405.1
ZC4	*Caldibacillus debilis*	91,47	0,26	GCF_001587535.1 GCF_000383875.1 GCF_003627895.1
ZC4	*Caldicoprobacter oshimai*	92,72	6,72	GCF_000526435.1
ZC4	*Clostridium cellulosi*	80,35	2,17	GCF_000953215.1
ZC4	*Mycobacterium hassiacum*	85,03	4,7	GCF_900603025.1
ZC4	*Planifilum fulgidum*	97,69	0,51	GCF_900113175.1
ZC4	*Pseudomonas thermotolerans*	94,23	3,52	GCF_000364625.1 GCF_000513835.1
ZC4	*Rhodothermus marinus*	99,44	0	GCA_009936255.1 GCA_009936275.1 GCF_000024845.1 GCF_000224745.1
ZC4	*Sphaerobacter thermophilus*	95,33	0,93	GCF_000024985.1
ZC4	*Thermobifida fusca*	99,45	0,55	GCF_015034585.1 GCF_000012405.1
ZC4	*Thermobispora bispora*	97,53	0,79	GCF_000092645.1
ZC4	*Thermocrispum agreste*	83,24	3,51	GCF_000427905.1
SRA	*Desulfurellaceae bacterium*	71,30	4,08	GCA_002238415.1
SRA	*Truepera sp*	64,36	3,25	GCA_002239005.1

Capítulo 7

Resultados da análise dos dados

7.1 Resultados gerais de execução do MAGset

Executamos MAGset em todos os conjuntos de dados descritos no Capítulo 6. Os resultados encontram-se na Tabela 7.1.

No total, foram 36 execuções separadas do MAGset utilizando os MAGs originais, e 34 execuções separadas do MAGset utilizando os MAGs remontados. Na Tabela 7.1 mostramos apenas os resultados em termos de RGIs, mas para aqueles genomas denotados por asterisco (*) também foram gerados os resultados de anotações com COG e CAZy e os resultados do pangenoma das comparações realizadas. Na página web https://projetos.lbi.iq.usp.br/magset/resultados-dissertacao/ (para acesso, utilizar as seguintes credenciais: usuário: *magset-user* senha: *results2021*) é possível navegar em todos os resultados em HTML obtidos com o MAGset neste trabalho.

7.2 Resultados relativos à remontagem de MAGs

A Tabela 7.1 também mostra o número de RGIs negativas que foram encontradas nas amostras, através da execução do módulo MAGcheck. Todos os MAGs com pelo menos uma RGI negativa encontrada nos *reads* das amostras foi remontado (34 das 36 comparações), e o resultado da remontagem está disponível na Tabela 7.2, que informa o ganho/perda relativo ao MAG original nos seguintes aspectos: tamanho, completude, contaminação, quantidade de RGIs negativas, quantidade de RGIs encontradas pelo MAGcheck e quantidade de genes identificados. As informações relativas ao tamanho, completude, contaminação e genes identificados foram obtidas utilizando o CheckM, e a quantidade de RGIs foi obtida utilizando o MAGset.

Todos os valores detalhados das comparações entre a versão original do MAG e sua versão remontada estão disponibilizados no Apêndice B.

Tabela 7.1: *Quantidade de RGIs encontradas por MAG analisado*

Origem	Espécie	RGIs negativas				RGIs positivas	
		Qtde	Tamanho total (pb)	Encontradas pelo MAGcheck	% das RGIs encontradas	Qtde	Tamanho total (pb)
Artigo FT	*Bacteroides uniformis*	**68**	**1.738.894**	**29**	42,65	**44**	**546.741**
Artigo FT	*Bacteroides vulgatus*	**95**	**1.501.981**	**52**	54,74	**128**	**1.189.764**
Artigo GI	*Clostridium baratii*	**34**	**673.847**	**0**	0,00	**16**	**216.513**
Artigo GI	*Enterococcus faecalis*	**20**	**566.124**	**0**	0,00	**15**	**213.077**
Artigo GI	*Enterococcus faecium*	**47**	**778.775**	**27**	57,45	**1**	**5.089**
Artigo HS	*Ardenticatena maritima*	**69**	**670.140**	**64**	92,75	**11**	**95.573**
Artigo HS	*Chloroflexus aurantiacus*	**50**	**526.461**	**46**	92,00	**4**	**29.291**
Artigo HS	*Thermosynechococcus sp*	**12**	**109.810**	**12**	100,00	**0**	**0**
Artigo PP	*Marivivens sp*	**19**	**696.613**	**6**	31,58	**3**	**42.889**
Artigo PP	*Ruegeria pomeroyi*	**7**	**46.167**	**7**	100,00	**1**	**6.765**
Artigo SG	*Enterococcus faecalis* *	**27**	**583.812**	**4**	14,81	**21**	**211.604**
Artigo SG	*Enterococcus faecium* *	**37**	**690.107**	**5**	13,51	**15**	**143.116**
Bugios	*Prevotella sp*	**32**	**396.439**	**1**	3,13	**60**	**594.557**
Bugios	*Treponema berlinense*	**27**	**364.754**	**2**	7,41	**15**	**222.923**
Bugios	*Treponema succinifaciens*	**36**	**829.801**	**6**	16,67	**16**	**138.822**
Composto Maduro	*Mycolicibacterium thermoresistibile* *	**64**	**939.991**	**46**	71,88	**3**	**24.844**
Composto Maduro	*Novibacillus thermophilus*	**109**	**1.629.372**	**97**	88,99	**1**	**7.982**
Composto Maduro	*Planifilum fulgidum*	**112**	**1.017.830**	**102**	91,07	**0**	**0**
Composto Maduro	*Thermocrispum municipale* *	**20**	**392.384**	**12**	60,00	**22**	**284.560**
Inóculo	*Sphaerobacter thermophilus*	**128**	**1.180.956**	**107**	83,59	**10**	**64.421**
Lago	*Limnohabitans sp*	**26**	**442.506**	**14**	53,85	**12**	**89.237**
ZC3	*Planifilum fulgidum* *	**19**	**325.864**	**13**	68,42	**5**	**36.182**
ZC3	*Thermobifida fusca* *	**20**	**227.414**	**11**	55,00	**3**	**27.111**
ZC4	*Caldibacillus debilis* *	**85**	**1.116.582**	**53**	62,35	**0**	**0**
ZC4	*Caldicoprobacter oshimai* *	**40**	**502.009**	**26**	65,00	**20**	**226.507**
ZC4	*Clostridium cellulosi* *	**36**	**371.654**	**27**	75,00	**0**	**0**
ZC4	*Mycobacterium hassiacum* *	**37**	**437.636**	**28**	75,68	**3**	**27.770**
ZC4	*Planifilum fulgidum* *	**16**	**190.338**	**14**	87,50	**4**	**28.312**
ZC4	*Pseudomonas thermotolerans* *	**46**	**526.695**	**28**	60,87	**8**	**67.320**
ZC4	*Rhodothermus marinus* *	**106**	**1.465.051**	**42**	39,62	**2**	**12.499**
ZC4	*Sphaerobacter thermophilus* *	**24**	**193.294**	**19**	79,17	**13**	**106.057**
ZC4	*Thermobifida fusca* *	**19**	**220.690**	**14**	73,68	**3**	**25.572**
ZC4	*Thermobispora bispora* *	**10**	**155.491**	**10**	100,00	**3**	**27.008**
ZC4	*Thermocrispum agreste* *	**14**	**126.014**	**12**	85,71	**9**	**66.484**
SRA	*Desulfurellaceae bacterium*	**174**	**2.017.769**	**108**	62,07	**6**	**37.806**
SRA	*Truepera sp*	**100**	**1.220.936**	**52**	52,00	**5**	**34.009**

Tabela 7.2: *Resultado da remontagem dos MAGs*

Origem	Espécie	Tamanho	Completude	Contaminação	RGIs negativas	RGIs MAGcheck	Genes identificados
Artigo FT	*Bacteroides uniformis*	+16,44	+3,40	-1,11	-23,53	-82,76	+12,37
Artigo FT	*Bacteroides vulgatus*	+7,05	+4,86	-4,30	-38,95	-92,31	+2,11
Artigo GI	*Enterococcus faecium*	+10,51	-0,60	+1,43	-31,91	-62,96	+10,46
Artigo HS	*Ardenticatena maritima*	+14,78	+16,08	+0,33	-81,16	-90,63	+12,52
Artigo HS	*Chloroflexus aurantiacus*	+6,82	+3,77	-0,94	-84,00	-91,30	+5,17
Artigo HS	*Thermosynechococcus sp*	+4,25	+0,47	-0,94	-100,00	-100,00	-1,68
Artigo PP	*Marivivens sp*	+11,16	+1,49	+1,15	-5,26	-83,33	+10,14
Artigo PP	*Ruegeria pomeroyi*	+0,77	+0,66	-0,90	-57,14	-57,14	-1,86
Artigo SG	*Enterococcus faecalis*	+1,14	0,00	-0,05	-3,70	-75,00	+0,68
Artigo SG	*Enterococcus faecium*	+3,10	0,00	-0,81	+2,70	-100,00	+3,36
Bugios	*Prevotella sp*	+3,89	0,00	+0,05	-9,38	-100,00	+5,99
Bugios	*Treponema berlinense*	+5,18	-1,79	+4,55	-18,52	-100,00	+13,59
Bugios	*Treponema succinifaciens*	+4,20	0,00	-0,65	-30,56	-100,00	+3,45
Composto Maduro	*Mycolicibacterium thermoresistibile*	+9,47	+7,00	-0,49	-68,75	-95,65	+7,59
Composto Maduro	*Novibacillus thermophilus*	+53,44	+12,74	+0,77	-77,06	-98,97	+60,66
Composto Maduro	*Planifilum fulgidum*	+24,06	+13,78	-1,22	-81,25	-93,14	+18,00
Composto Maduro	*Thermocrispum municipale*	+9,47	+7,00	-0,49	-50,00	-83,33	+7,59
Inóculo	*Sphaerobacter thermophilus*	+24,91	+12,78	+1,93	-82,03	-96,26	+27,12
Lago	*Limnohabitans sp*	+7,46	+0,62	-0,58	-38,46	-92,86	+6,18
ZC3	*Planifilum fulgidum*	+7,64	+4,61	+0,77	-78,95	-100,00	+6,96
ZC3	*Thermobifida fusca*	+0,71	0,00	0,00	-75,00	-100,00	+0,66
ZC4	*Caldibacillus debilis*	+9,76	+0,12	+0,42	-50,59	-86,79	+11,29
ZC4	*Caldicoprobacter oshimai*	+1,20	+0,80	+0,16	-27,50	-34,62	-4,94
ZC4	*Clostridium cellulosi*	+10,70	+3,36	-1,28	-72,22	-92,59	+4,23
ZC4	*Mycobacterium hassiacum*	+2,15	+1,74	-2,73	-64,86	-92,86	-6,08
ZC4	*Planifilum fulgidum*	+4,45	0,00	0,00	-75,00	-85,71	+3,92
ZC4	*Pseudomonas thermotolerans*	+4,68	+1,10	+1,29	-54,35	-82,14	+5,42
ZC4	*Rhodothermus marinus*	+5,43	+0,56	0,00	-32,08	-85,71	+4,43
ZC4	*Sphaerobacter thermophilus*	+1,72	-0,94	-0,93	-75,00	-94,74	-2,21
ZC4	*Thermobifida fusca*	+0,55	0,00	-0,44	-73,68	-92,86	-2,87
ZC4	*Thermobispora bispora*	+3,80	+1,41	-0,26	-100,00	-100,00	+3,62
ZC4	*Thermocrispum agreste*	+0,57	+0,45	-0,80	-85,71	-91,67	-7,12
SRA	*Desulfurellaceae bacterium*	+41,89	+5,88	+2,53	-36,78	-65,74	+43,25
SRA	*Truepera sp*	+32,54	+9,53	+0,42	-66,00	-96,15	+30,10

Valores expressos em percentuais, representando os ganhos ou perdas de cada característica avaliada. Em verde estão os valores que o resultado foi melhor que o MAG original, e em vermelho os valores que o resultado foi pior que o MAG original. A coluna "RGI negativas" indica a variação percentual de RGIs negativas encontradas após a remontagem. A coluna "RGIs MAGcheck" indica a variação percentual das RGIs negativas encontradas pelo MAGcheck após a remontagem, em comparação com a quantidade encontrada pelo MAGcheck antes da remontagem.

7.3 Análises especiais

7.3.1 Genes de RNA ribossomal recuperados

Além dos resultados gerais de execução de MAGset, MAGcheck e remontagem, apresentados nas duas seções anteriores, nesta seção apresentamos resultados mais específicos, relativos à detecção de genes de RNA ribossomal nas amostras.

Escolhemos o MAG *Thermobifida fusca* do projeto ZC4 para uma análise mais detalhada, já que os dois genomas completos inseridos na comparação com o MAG contêm mais genes de RNA ribossomal que o MAG. A cepa YX contém 4 conjuntos de genes de RNA ribossomal completos (5S, 16S, e 23S) e a cepa UPMC_901 contém 3 conjuntos de genes de RNA ribossomal completos. Já o MAG original contém apenas um conjunto de genes de RNA ribossomal parcial, com apenas o gene 5S. Como é possível observar na Tabela B.121 a completude do MAG original já é muito alta (99%).

Ao utilizar a busca avançada de genes da ferramenta MAGset para verificar se existiam genes de RNA ribossomal nas RGIs negativas, foi possível identificar que os genes estavam em RGIs, e que as RGIs foram encontradas pelo MAGcheck nas amostras, como é possível observar na Figura 7.1.

Após a remontagem do MAG, foi realizada uma busca com *blastn* para verificar se uma quantidade maior de genes de RNA ribossomal foram incorporados no MAG remontado, e foi possível encontrar 1 gene 16S completo, 1 gene 23S parcial (faltando 141 de 3015 pb) e 3 genes 16S (2 completos e um parcial).

Além da busca via *blastn*, também foi feita a anotação dos genomas com o programa PGAP [22], e o mesmo resultado foi encontrado (Tabela 7.3).

Tabela 7.3: *Comparação de genes de RNA ribossomal - Thermobifida fusca - ZC4*

	genes completos			**genes incompletos**		
	16S	**23S**	**5S**	**16S**	**23S**	**5S**
ZC4RG21	0	0	0	0	0	1
ZC4RG21 remontado	1	0	2	0	1	1
Diferença	**+1**	-	**+2**	-	**+1**	-

Genes

Selected columns

Show 10 entries

Search:

Genome	Locus tag	Type	Product	Parent	GRI ID	Found by MAGcheck
UPMC_901	IM867_RS01095	rRNA	16S ribosomal RNA	NZ_CP063232	NGRI0005_04_UPMC_901	true
UPMC_901	IM867_RS03900	rRNA	16S ribosomal RNA	NZ_CP063232	NGRI0005_05_UPMC_901	true
UPMC_901	IM867_RS14560	rRNA	16S ribosomal RNA	NZ_CP063232	NGRI0005_06_UPMC_901	true
YX	TFU_RS10060	rRNA	16S ribosomal RNA	NC_007333	NGRI0005_01_YX	true
YX	TFU_RS11740	rRNA	16S ribosomal RNA	NC_007333	NGRI0006_01_YX	true
YX	TFU_RS12830	rRNA	16S ribosomal RNA	NC_007333	NGRI0005_02_YX	true
YX	TFU_RS15320	rRNA	16S ribosomal RNA	NC_007333	NGRI0005_03_YX	true

Showing 1 to 7 of 7 entries

Previous 1 Next

Export results to CSV

Figura 7.1: *Resultados obtidos com MAGset para a busca por genes de RNA ribossomal*

Para obter esses resultados, foram utilizados os seguintes filtros dentro do MAGset: *"found by MAGcheck equals true"* interseccionado com *"product contains 16S"*

7.3.2 Genes de fagos no MAG *Pseudomonas thermotolerans* (ZC4)

Escolhemos também o MAG *Pseudomonas thermotolerans* do projeto ZC4 para analisar as RGIs positivas encontradas, já que essa espécie é de interesse de outros projetos e estudos do nosso laboratório, por sua capacidade de degradação de matéria orgânica. Apesar de não encontrar genes relacionados a degradação, listamos os genes encontrados na RGI positiva PGRI0041 na Tabela 7.4 que identificamos características interessantes para análise. Calculamos o percentual do conteúdo GC da sequência desse RGI, resultando em 64,85%, contra 67,2% do MAG.

Tabela 7.4: *Genes da RGI positiva PGRI0041 do MAG Pseudomonas thermotolerans - ZC4*

Locus tag	Tipo	Produto	Identificador da proteína
C0P69_14685	CDS	*glycoprotein*	PRJNA317502:C0P69_14685
C0P69_14690	CDS	*Ig domain-containing protein*	PRJNA317502:C0P69_14690
C0P69_14695	CDS	*hypothetical protein*	PRJNA317502:C0P69_14695
C0P69_14700	CDS	*hypothetical protein*	PRJNA317502:C0P69_14700
C0P69_14705	CDS	*hypothetical protein*	PRJNA317502:C0P69_14705
C0P69_14710	CDS	*hypothetical protein*	PRJNA317502:C0P69_14710
C0P69_14715	CDS	*hypothetical protein*	PRJNA317502:C0P69_14715
C0P69_14720	CDS	*hypothetical protein*	PRJNA317502:C0P69_14720
C0P69_14725	CDS	*hypothetical protein*	PRJNA317502:C0P69_14725
C0P69_14730	CDS	*hypothetical protein*	PRJNA317502:C0P69_14730
C0P69_14735	CDS	*phage head morphogenesis protein*	PRJNA317502:C0P69_14735
C0P69_14740	CDS	*DUF1073 domain-containing protein*	PRJNA317502:C0P69_14740
C0P69_14745	CDS	*aldehyde dehydrogenase*	PRJNA317502:C0P69_14745
C0P69_14750	CDS	*terminase*	PRJNA317502:C0P69_14750
C0P69_14755	CDS	*terminase small subunit*	PRJNA317502:C0P69_14755
C0P69_14760	CDS	*phage holin, lambda family*	PRJNA317502:C0P69_14760
C0P69_14765	CDS	*hypothetical protein*	PRJNA317502:C0P69_14765
C0P69_14770	CDS	*hypothetical protein*	PRJNA317502:C0P69_14770
C0P69_14775	CDS	*endodeoxyribonuclease RusA*	PRJNA317502:C0P69_14775
C0P69_14780	CDS	*hypothetical protein*	PRJNA317502:C0P69_14780
C0P69_14785	CDS	*DUF1364 domain-containing protein*	PRJNA317502:C0P69_14785
C0P69_14790	CDS	*hypothetical protein*	PRJNA317502:C0P69_14790
C0P69_14795	CDS	*TraR/DksA family transcriptional regulator*	PRJNA317502:C0P69_14795
C0P69_14800	CDS	*hypothetical protein*	PRJNA317502:C0P69_14800
C0P69_14805	CDS	*hypothetical protein*	PRJNA317502:C0P69_14805
C0P69_14810	CDS	*NYN domain-containing protein*	

Capítulo 8

Discussão dos resultados de análise dos dados

Conforme mencionado no capítulo 3.1, embora a ferramenta MAGset permita análise e comparação de MAGs e genomas de referência, tal funcionalidade é limitada e não se compara com funcionalidades deste tipo providas por plataformas tais como IMG e PATRIC.

Por este motivo, concentramos nossa atenção neste capítulo na funcionalidade do MAGset relativa às regiões genômicas de interesse.

8.1 RGIs negativas

Uma das funcionalidades que acreditamos ser mais valiosa no MAGset é sua capacidade de detectar RGIs negativas. Estas, como explicado anteriormente, são regiões presentes no genoma de referência mas ausentes no MAG de interesse.

A quantidade de RGIs negativas foi considerável em grande parte dos genomas comparados, em alguns casos com mais de 100 RGIs, como nos MAGs das espécies *Novibacillus thermophilus* (B.62) com 109 RGIs, *Planifilum fulgidum* (B.66) com 112 RGIs, *Sphaerobacter thermophilus* (B.74) com 128 RGIs, *Rhodothermus marinus* (B.114) com 106 RGIs e *Desulfurellaceae bacterium* (B.134) com 174 RGIs. É importante salientar que algumas comparações foram feitas com mais de um genoma de referência, o que pode aumentar a quantidade de RGIs negativas encontradas, já que a mesma região pode existir em mais de um genoma de referência, e assim ser reportada mais de uma vez.

Entre as comparações realizadas, os MAGs em que encontramos menos RGIs negativas foram *Ruegeria pomeroyi* (B.34) com 7 RGIs, *Thermobispora bispora* (B.126) com 10 RGIs e *Thermocrispum agreste* (B.130) com 14 RGIs. Esses genomas com poucas RGIs também são muito similares às referências quando olhamos para o ANI (99,97%, 99,75% e 99,72%, respectivamente). Se a montagem está com uma boa qualidade e são muito parecidos com os genomas de referência, é esperado que poucas RGIs negativas sejam encontradas, como foi o caso destes genomas.

Quando olhamos para o tamanho total das RGIs negativas encontradas nas comparações realizadas, o maior resultado encontrado foi 2.017.769 pb, no MAG *Desulfurellaceae bacterium* (B.134). É importante relembrar que esse MAG foi comparado com uma referência obtida de um ambiente metagenômico, ou seja, também é um MAG, o que pode ter gerado contaminação, e essa contaminação da referência irá aparecer nos resultados do MAGset como RGIs negativas encontradas. Por esse motivo, a comparação com genomas de referências obtidos de isolados é sempre mais adequada e confiável. No caso desse MAG, infelizmente não existiam genomas de referência a partir de isolados, por isso a comparação

foi feita com outro MAG, já publicado anteriormente. Além disso, o resultado ANI entre os dois genomas foi quase no limite da definição de genomas da mesma espécie (95,67%), o que também aumenta as chances de se obter uma grande quantidade de RGIs, já que os genomas são mais distantes, apesar de ainda serem da mesma espécie.

Uma vez que foram encontradas as RGIs negativas, uma questão relevante é tentar identificar quais são RGIs negativas verdadeiras (regiões que realmente não existem no genoma do organismo correspondente ao MAG, e existem na referência) ou RGIs negativas falsas (regiões que foram identificadas como RGIs negativas, mas não estão nos MAGs somente por uma falha no processo de montagem). Como exemplo de RGIs negativas verdadeiras temos as ilhas genômicas (2.11), que foram adquiridas após a divergência das cepas.

Já RGIs negativas falsas podem ter ocorrido por uma baixa/nenhuma cobertura de uma parte do genoma ou um erro no processo de montagem do MAG. É exatamente esta observação que nos motivou a criar o módulo MAGcheck, que verifica se as RGIs negativas estão nos *reads* das amostras que deram origem ao MAG. Ao encontrar as RGIs nas amostras, é possível tentar utilizar esses *reads* para melhorar o MAG original, como discutiremos mais abaixo, no item 8.3. Mas o fato de encontrar as RGIs nas amostras já é um forte indicativo que aquela região é uma RGI negativa falsa, e isso deve ser considerado nas análises de diferenças entre os genomas.

Com as RGIs negativas verdadeiras a situação é mais delicada, já que mesmo que o MAGcheck não localize essa região nas amostras, isso não é garantia que o genoma realmente não contenha essa RGI, pois limitações no processo de obtenção das amostras podem causar perdas de partes do genoma. Ainda que em MAGs sempre paire a dúvida sobre esses pontos, podemos considerar que não encontrar a RGI nas amostras é mais uma evidência que ela é uma RGI negativa verdadeira, dentro das limitações inerentes aos MAGs.

Ainda que as RGIs negativas não necessariamente representem todas as partes faltantes no MAG, já que neste trabalho consideramos apenas as diferenças maiores que 5 mil pb, observamos casos em que todas as RGIs negativas foram encontradas nas amostras. Um exemplo é o MAG da espécie *Thermobispora bispora* da composteira ZC4, para o qual as 10 RGIs negativas detectadas por MAGcheck foram encontradas nas amostras. O mesmo ocorreu com o MAG da espécie *Thermosynechococcus sp* do artigo HT e no MAG *Ruegeria pomeroyi* do artigo PP.

Essas partes desconsideradas podem influenciar na análise dos genomas analisados, bem como influenciar a análise de funções que os genomas têm no ambiente, pois os genes contidos nas RGIs negativas falsas são ignorados.

Mesmo que para muitos casos a quantidade de RGIs negativas encontradas nas amostras seja considerável, também existem casos analisados com nenhuma ou poucas RGIs encontradas, como no MAG da espécie *Clostridium baratii* e da espécie *Enterococcus faecalis*, ambos do artigo GI, com nenhuma RGI negativa encontrada nas amostras.

É importante salientar que o tamanho mínimo que utilizamos para RGIs foi de 5.000 pb. Ao utilizar valores menores, corre-se o risco de obter muitas RGIs sem valor biológico, criando muito ruído ao avaliar as RGIs encontradas, e ao utilizar RGIs muito grandes, é possível que diversas regiões de importância sejam perdidas na comparação. De qualquer forma, esse valor é configurável pelo usuário do programa, e vai depender do objetivo do pesquisador, se deseja um refinamento mais detalhado na comparação (RGIs menores) ou se busca grandes diferenças entre os genomas (RGIs maiores).

A grande quantidade de RGIs negativas encontradas nas amostras é mais uma confirmação de que no processo de montagem de MAGs diversas partes são desconsideradas pelos processos de montagem de genomas de dados metagenômicos que consideramos neste trabalho. Mesmo que os trabalhos tenham utilizados diferentes softwares para a montagem, como

descrito na Tabela A.2, diversas RGIs foram encontradas em todos os casos.

8.2 RGIs positivas

Como já explicado anteriormente, RGIs positivas são regiões que não existem em nenhuma referência, mas existem no MAG. Essa regiões podem ser RGIs positivas verdadeiras (adquiridas posteriormente à divergência das cepas ou regiões perdidas no processo de evolução do genoma de referência, por exemplo), ou podem ser RGIs positivas falsas, inseridas no MAG por erro no processo de montagem do genoma, o que também chamamos de contaminação. Infelizmente as ferramentas disponíveis para se verificar se a região é falsa ou verdadeira são limitadas, principalmente se a RGI não está contida em um *contig* com outras partes.

Caso a RGI esteja contida em um *contig* que contenha partes que são compartilhadas com genomas de referência, temos uma evidência mais forte que a mesma é verdadeira, e essa pode ser mais uma estratégia de verificação; no entanto não podemos automaticamente concluir o contrário: Se uma RGI positiva representa um *contig* inteiro, ainda assim ela pode ser verdadeira também, já que para fazer parte do MAG, passou por alguns filtros de similaridade com o restante do MAG para ser incorporada. Análises mais finas serão necessárias caso o pesquisador queira validar em detalhes a RGI positiva, como verificar quais genes fazem parte da RGI e se tais genes tem relação forte com o MAG.

Em geral, a quantidade de RGIs positivas encontradas foi menor que a quantidade de RGIs negativas, o que é esperado, já que os MAGs são mais incompletos que os genomas de referência (quando estes são genomas completos de isolados). As exceções foram o MAG *Bacteroides vulgatus* (B.6) do artigo FT, o MAG *Prevotella sp* (B.46) dos dados de Bugios do projeto Metazoo e o MAG *Thermocrispum municipale* (B.70) dos dados do Composto Maduro do projeto Metazoo.

No caso do MAG *Bacteroides vulgatus*, é possível verificar que sua contaminação estava alta (8,56%) e o seu tamanho é maior que o tamanho do genoma de referência (5,43 Mpb versus 5,16 Mpb, respectivamente), ainda que sua completude seja bem menor (79,53% versus 99,25%), conforme descrito na Tabela B.5, o que indica que possivelmente algumas das RGIs positivas são contaminações. Já no caso do MAG *Prevotella sp*, a contaminação está em apenas 1,11% (Tabela B.45), o que indica que o CheckM não identificou essa grande quantidade de RGIs positivas como contaminação. Isso não é garantia que as RGIs não são contaminação, já que o CheckM muitas vezes não consegue identificar toda a contaminação, como pontuado pelos autores em [35], sendo necessário mais investigações. Uma possível forma de investigar este ponto é analisar os genes dentro dessas regiões, para se ter mais confiança se a região é uma contaminação ou realmente uma RGI positiva verdadeira. Já no caso do MAG *Thermocrispum municipale*, tivemos quase a mesma quantidade de RGIs negativas e positivas (20 e 22, respectivamente). Já quando olhamos o tamanho total das RGIs negativas e positivas, percebemos que o tamanho das RGIs negativas é consideravelmente maior (392.384 pb para negativas e 284.560 pb para positivas), ou seja, existem mais RGIs positivas, mas quando olhamos para o tamanho total, vemos que essa situação só ocorre porque as RGIs positivas estão mais fragmentadas.

É importante mencionar que existe uma assimetria entre RGIs negativas verdadeiras e possíveis ilhas genômicas do genoma correspondente ao MAG de interesse. No primeiro caso, uma inspeção da RGI no genoma de referência em muitos casos pode nos dar evidências de que a RGI é uma ilha genômica, devido ao seu conteúdo gênico e outras características. O importante aqui é o fato de que o trecho genômico que motivou sua designação como RGI negativa nos é acessível. No caso de uma ilha genômica no genoma correspondente ao

MAG de interesse, existe uma chance de que tal ilha não tenha sido incorporada ao MAG, justamente por ser uma ilha, e portanto com características (principalmente de composição de nucleotídeos) marcadamente diferentes do resto do genoma. Tais ilhas serão invisíveis, pois não vão nem sequer aparecer como RGIs positivas (o que seria o desejado). A nosso ver isto mostra uma das grandes limitações do processo de recuperação de MAGs, e que necessita de pesquisas adicionais para ser satisfatoriamente solucionada.

Fizemos uma análise amostral de RGIs positivas para tentar verificar sua origem. Os genes encontrados em uma RGI positiva do MAG *Pseudomonas thermotolerans* do projeto ZC4 contém a descrição de genes geralmente identificados em fagos (listados na Tabela 7.4), sugerindo que essa região foi obtida via transferência horizontal. Esse MAG foi analisado em mais detalhes em busca de genes de degradação de matéria orgânica, que não foram encontrados nas RGIs positivas. Em contrapartida, esta RGI com genes de fagos chamou a atenção, pois a RGI representa um *contig* inteiro no MAG, e como já discutido neste trabalho, geralmente elementos móveis não são identificados em MAGs. Ao analisar o percentual do conteúdo GC da RGI com o percentual do conteúdo GC do MAG, temos 64,85% para a RGI, e 67,2% para o MAG, uma pequena diferença mas sugestiva de que se trata de uma RGI positiva verdadeira. Uma hipótese para sua presença é que ela foi resultado da incorporação de um fago, transformando-se em profago e, com o passar do tempo, "aclimatando-se" ao genoma do hospedeiro em termos de sua composição. Este é um fenômeno bem conhecido [51]. Quanto mais gerações passam após a aquisição da região incorporada, com o acumulo de mutações, mais similar ao restante do genoma a região se torna. Se esta hipótese for correta, podemos formular uma segunda hipótese: de que regiões adquiridas por transferência lateral em genomas de MAGs só farão parte do MAG caso pelo menos uma de duas condições sejam satisfeitas: 1) o organismo do qual proveio o trecho tinha originalmente uma composição nucleotídica semelhante ao do MAG; 2) o evento de transferência lateral é antigo, portanto houve tempo para a aclimatação referida. Um corolário desta hipótese é que regiões adquiridas lateral e recentemente, com composição de bases substancialmente diferente do MAG, não farão parte do MAG, sendo difícil sua localização nas amostras (ilhas genômicas invisíveis, conforme discutido no parágrafo anterior).

8.3 Resultados obtidos após remontagem dos MAGs

A remontagem de MAGs trouxe melhorias no tamanho dos MAGs em todos os casos estudados. Os MAGs que tiveram poucas RGIs negativas encontradas em geral tiveram um menor incremento em seu tamanho final, como esperado, já que existiam menos regiões encontradas nas amostras, e consequentemente, menos regiões a serem incorporadas no MAG.

A completude dos genomas foi melhorada em diversos casos, com destaque para o MAG da espécie *Ardenticatena maritima* do artigo HS, que originalmente tinha 70% e após a remontagem atingiu 86% de completude, com incremento mínimo na contaminação, conforme é possível verificar na Tabela B.17.

A quantidade de genes identificados nos genomas também teve um incremento considerável (em 27 dos 34 MAGs remontados), o que pode auxiliar o estudo e compreensão dos ambientes onde os MAGs foram originalmente obtidos.

É importante salientar que uma RGI negativa encontrada nos *reads* das amostras não necessariamente será incorporada ao MAG após realizar a remontagem. Isso ocorre porque muitas vezes a cobertura da RGI é muito baixa e com *gaps* (buracos), o que impede o montador de conseguir obter um *contig* dessas regiões.

Após a remontagem, alguns MAGs ficaram sem nenhuma RGI negativa, ou seja, todas as RGIs negativas que existiam antes estavam nos *reads* das amostras e foram corretamente

remontados. Isso ocorreu no MAG *Thermosynechococcus sp* do artigo HS, que inicialmente tinha 12 RGIs negativas, todas encontradas pelo MAGcheck (Tabela B.26), e após a remontagem, nenhuma RGI negativa foi encontrada (Tabela B.27). O mesmo ocorreu com o MAG *Thermobispora bispora*, dos dados da composteira ZC4 do projeto Metazoo (Tabela B.127).

Até mesmo RNAs ribossomais foi possível recuperar após a remontagem de um dos MAGs, como mostrado na seção 7.3.1. Essa é mais uma evidência da possibilidade de encontrar partes "perdidas" do MAG nas amostras, com a possibilidade de incrementar a qualidade final do MAG.

Ainda que não tenha sido possível recuperar todos os RNAs ribossomais, como seria desejado, já que a montagem de MAGs tem essa limitação já conhecida [34], a inclusão de mais alguns genes dessa relevância no MAG demonstra a importância da revisão da montagem e os ganhos que isso pode trazer.

Estes resultados indicam que existe espaço para melhoria nos processos de recuperação de MAGs.

8.4 Auto-avaliação sobre os principais resultados conseguidos

Nossa auto-avaliação dos resultados mais relevantes obtidos é a seguinte:

- MAGset nos permitiu encontrar dezenas de RGIs (positivas e negativas) nos MAGs testados de forma automática e eficiente;
- MAGset é uma ferramenta que oferece uma interface simples e intuitiva para análise do conteúdo gênico das RGIs encontradas;
- A investigação das RGIs negativas nas amostras permitiu melhorias significativas das montagens na grande maioria dos MAGs analisados; este resultado por sua vez indica que ainda há um bom espaço para que melhores softwares de *binning* sejam desenvolvidos;
- Uma limitação de MAGset é que ele está restrito a lidar com MAGs que tenham sido classificados ao nível de espécie. Tendo em vista o último resultado elencado, gostaríamos de generalizar o processo de busca de RGIs negativas nas amostras para MAGs sem espécie conhecida. No entanto, fazer isto ficou além do escopo deste trabalho.

Capítulo 9

Conclusão

9.1 Contribuições do trabalho

O software proposto neste trabalho permite identificar particulares entre MAGs e referências da mesma espécie, além de validar se as particularidades que o MAG não contém (chamadas de RGIs negativas) estão nos *reads* das amostras que deram origem ao MAG, trazendo mais informações ao usuário da ferramenta sobre o MAG e permitindo um melhor entendimento sobre as partes que o MAG não contém, que podem ter sido causadas por um erro no processo de montagem.

O software foi disponibilizado gratuitamente no GitHub, tanto para ser utilizado em outros trabalhos, com instruções para instalação e tutoriais, como para ser estendido/modificado por eventuais interessados em evoluir o software.

Adicionalmente, foi demonstrado que em diversos casos é possível utilizar os resultados da ferramenta MAGset para reconstruir MAGs, melhorando consideravelmente o tamanho e completude do MAG reconstruído em diversos casos.

9.2 Trabalhos futuros

Trabalhos futuros podem estender o software MAGset para suportar diversas novas funcionalidades:

9.2.1 Realizar mais tipos de anotações

Hoje o suporte das anotações se restringe às anotações COG e CAZy, mas outros tipos de anotação podem enriquecer os resultados apresentados;

9.2.2 Suportar genomas anotados em outros formatos

Hoje somente genomas anotados no formado GenBank são suportados (.gbff), no entanto genomas anotados por outras ferramentas, como o PROKKA [21] são muito comuns também, e cada vez mais utilizados.

9.2.3 Validar o MAGset para *reads* longos

O software MAGset foi validado apenas com *reads* curtos, não sendo possível avaliar se é possível obter resultados relevantes com *reads* longos, e como estes estão se tornando cada mais comuns, seria uma funcionalidade interessante para disponibilizar.

9.2.4 Inserir o software CheckM no *pipeline*

Como o CheckM [19] é uma ferramenta muito utilizada para avaliar a completude e contaminação dos genomas, e frequentemente faz parte da análise em MAGs, integrá-la ao MAGset vai trazer mais praticidade aos usuários se esses resultados já estiverem disponíveis junto com as outras informações reportadas.

9.2.5 Integrar ao MAGset a remontagem dos MAGs

Assim como feito de forma manual nesse trabalho, a remontagem dos MAGs pode ser automatizada e inserida como opção do software. O aumento de complexidade é considerável, já que diversos novos parâmetros (tamanho mínimo de *contig* aceito, por exemplo) e novos passos deveriam ser suportados (programas para o controle de qualidade dos *reads*, como o TrimGalore: https://www.bioinformatics.babraham.ac.uk/, remontagem utilizando diferentes softwares, como o Spades [13] e o MEGAHIT [14]). A complexidade de execução também irá aumentar, já que o uso de recursos como memória e tempo de processamento é muito variável de acordo com o tamanho dos dados, podendo demorar muitas horas para finalizar o processo. Como essa melhoria aumentaria demais o escopo deste trabalho, ela não foi realizada.

9.2.6 Melhorar a montagem de MAGs na parte de ilhas genômicas

Ilhas genômicas são um grande desafio para serem recuperadas quando estamos trabalhando com dados metagenômicos. Como essas regiões em geral não seguem o mesmo padrão de conteúdo do restante do genoma, os processos de montagem não conseguem identificar essas regiões e inserir dentro dos MAGs obtidos. Caso a ilha genômica apareça como uma RGI negativa, e esteja na amostra, é possível recuperá-la seguindo o processo descrito neste trabalho. Mas se a ilha genômica for específica do genoma no ambiente metagenômico em relação ao genoma de referência, ela não existirá no genoma de referência e por consequência não será recuperada. Uma forma de obter ilhas genômicas nos MAGs sugerida pelo autor em [52] é a utilização de tecnologias que obtém *reads* longos, inclusive sendo possível mesclar montagens de *reads* longos com *reads* curtos, além de utilizar tecnologias mais atuais, como o *single-cell*, que consegue extrair DNA de apenas uma célula. No entanto, para projetos com apenas *reads* curtos, como os projetos aqui analisados, não temos uma sugestão neste momento de como fazer isso, e este é um assunto importante e pode ser endereçado em trabalhos futuros.

Apêndice A

Amostras metagenômicas e ferramentas utilizadas

Tabela A.1: *Lista das amostras metagenômicas utilizadas*

Origem	**Qtde de amostras**	**Tamanho total (GB)**	**Tipo**	**Origem**	**Código das amostras utilizadas**	**Observações**
Artigo FT	3	39,4	*Paired-end*, Illumina	PRJNA510036	SRR8317179 SRR8317180 SRR8317181	Amostras do doador, relacionadas aos 2 MAGs que analisamos.
Artigo GI	2	140	*Paired-end*, Illumina	PRJNA294605	SRR3466404 SRR6257383	A relação entre qual genoma utiliza qual amostra foi obtida pelo arquivo aax5727_Table_S3.xlsx
Artigo HT	4	206	*Paired-end*, Illumina	PRJNA392119	SRR7905022 SRR7905023 SRR7905024 SRR7905025	Para melhorar os resultados, os dados metagenômicos utilizados foram filtrados com o software trim_galore antes de serem utilizados
Artigo PP	6	93,7	*Single-end*, Illumina	PRJNA553557	SRR11434620 SRR11434621 SRR11434623 SRR11434624 SRR11434626 SRR11434631	Para melhorar os resultados, os dados metagenômicos utilizados foram filtrados com o software trim_galore antes de serem utilizados
Artigo SG	3	70,6	*Paired-end*, Illumina	PRJEB18265	Hp_5_S19 Hp_6_S28 Hp_7_S14	Os códigos das amostras estão em outro artigo referenciado: https://doi.org/10.1016/j.dib.2017.01.007, item "1. DATA", "internal patient id"= 3
Bugios	13	7,5	*Paired-end*, Illumina	Dados internos do laboratório	Todas as amostras de cativeiro	Para melhor resultados, os dados metagenômicos utilizados são os *reads* já corrigidos pelo montador Spades
Composto maduro	2	3,9	*Paired-end*, Illumina	Dados internos do laboratório	ZCM1, ZCM2	Para melhor resultados, os dados metagenômicos utilizados são os *reads* já corrigidos pelo montador Spades
Inóculo	5	4,9	*Paired-end*, Illumina	Dados internos do laboratório	ZCI1, ZCI2, ZCI3, ZCI4, ZCI5	Para melhor resultados, os dados metagenômicos utilizados são os *reads* já corrigidos pelo montador Spades
Lago	26	41	*Paired-end*, Illumina	Dados internos do laboratório	Todas as amostras coletadas	Para melhorar os resultados, os dados metagenômicos utilizados foram filtrados com o software trim_galore antes de serem utilizados
ZC3	5	8,5	*Paired-end*, Illumina	Dados internos do laboratório	01, 30, 64, 78, 99	Para melhor resultados, os dados metagenômicos utilizados são os *reads* já corrigidos pelo montador Spades
ZC4	10	51	*Paired-end*, Illumina	Dados internos do laboratório	00, 01, 03, 07, 15, 30, 64, 67, 78, 99	Para melhor resultados, os dados metagenômicos utilizados são os *reads* já corrigidos pelo montador Spades
SRA	35	147	*Paired-end*, Illumina	Dados internos do laboratório	Todas as amostras coletadas	Para melhor resultados, os dados metagenômicos utilizados são os *reads* já descontaminados

Tabela A.2: *Softwares para obtenção de MAGs utilizados pelos artigos/projetos*

Origem	**Pipeline**	**Montagem**	**Binning**
Artigo FT	MetaWRAP	MEGAHIT	CONCOCT, MetaBat2, MaxBin2
Artigo GI	*Pipeline* próprio	IDBA-UD	abawaca, CONCOCT, MaxBin2
Artigo HS	*Pipeline* próprio	MEGAHIT	CONCOCT, MetaBat2, MaxBin2
Artigo PP	Anvi'o	MEGAHIT	CONCOT
Artigo SG	MetaWRAP	MEGAHIT	CONCOCT, MetaBat2, MaxBin2
Bugios	MetaWRAP	Spades	CONCOCT, MetaBat2, MaxBin2
Composto maduro	MetaWRAP	Spades	CONCOCT, MetaBat2, MaxBin2
Inóculo	MetaWRAP	Spades	CONCOCT, MetaBat2, MaxBin2
Lago	MetaWRAP	Spades	CONCOCT, MetaBat2, MaxBin2
ZC3	*Pipeline* próprio	Spades	Metabat2
ZC4	*Pipeline* próprio	Spades	Metabat2
SRA	MetaWRAP	MEGAHIT	CONCOCT, MetaBat2, MaxBin2

Apêndice B

Resultados detalhados das comparações realizadas

As informações: tamanho, completude, contaminação e quantidade de genes foram obtidas utilizando o software CheckM [19].

B.1 Comparações do artigo FT

B.1.1 *Bacteroides uniformis*

Tabela B.1: *Artigo FT: Bacteroides uniformis - Dados dos genomas comparados*

	Tipo	**Tamanho (Mpb)**	**Completude (%)**	**Contaminação (%)**	**Qtde de genes**
Donor.bin.12	MAG	3,77	92,55	1,73	3.378
Donor.bin.12 remontado	MAG	4,39	95,95	0,62	3.796
NBRC_113350 (GCF_006742345.1)	Referência	4,99	99,26	0,87	4.244

Tabela B.2: *Artigo FT: Bacteroides uniformis - MAG versus referência*

Código	**ANI** (%)	**RGIs detectadas**		**RGIs detectadas pelo MAGcheck**	
		Qtde	**Tamanho total (pb)**	**Qtde**	**Tamanho total (pb)**
Donor.bin.12	-	44	546.741	-	-
NBRC_113350	98,64	68	1.738.894	29	633.016

Tabela B.3: *Artigo FT: Bacteroides uniformis - MAG remontado versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
Donor.bin.12 remontado	-	41	555.220	-	-
NBRC_113350	98,59	52	1.008.680	5	154.694

Tabela B.4: *Artigo FT: Bacteroides uniformis: MAG versus MAG remontado*

	Tamanho (Mpb)	Completude (%)	Contaminação (%)	RGIs negativas	RGIs MAGcheck	Genes identificados
Donor.bin.12	3,77	92,55	1,73	68	29	3.378
Donor.bin.12 remontado	4,39	95,95	0,62	52	5	3.796
Diferença (%)	**+16,44**	**+3,40**	**-1,11**	**-23,53**	**-82,76**	**+12,37**

B.1.2 *Bacteroides vulgatus*

Tabela B.5: *Artigo FT: Bacteroides vulgatus - Dados dos genomas comparados*

	Tipo	Tamanho (Mpb)	Completude (%)	Contaminação (%)	Qtde de genes
Donor.bin.47	MAG	5,43	79,53	8,56	5.506
Donor.bin.47 remontado	MAG	5,81	84,39	4,26	5.622
ATCC_8482 (GCF_000012825.1)	Referência	5,16	99,25	0,94	4.218

Tabela B.6: *Artigo FT: Bacteroides vulgatus - MAG versus referência*

Código	**ANI** (%)	**RGIs detectadas**		**RGIs detectadas pelo MAGcheck**	
		Qtde	**Tamanho total (pb)**	**Qtde**	**Tamanho total (pb)**
Donor.bin.47	-	128	1.189.764	-	-
ATCC_8482	96,28	95	1.501.981	52	496.935

Tabela B.7: *Artigo FT: Bacteroides vulgatus - MAG remontado versus referência*

Código	**ANI** (%)	**RGIs detectadas**		**RGIs detectadas pelo MAGcheck**	
		Qtde	**Tamanho total (pb)**	**Qtde**	**Tamanho total (pb)**
Donor.bin.47 remontado	-	124	1.237.483	-	-
ATCC_8482	96,20	58	744.261	4	31.603

Tabela B.8: *Artigo FT: Bacteroides vulgatus: MAG versus MAG remontado*

	Tamanho (Mbp)	**Completude (%)**	**Contaminação (%)**	**RGIs negativas**	**RGIs MAGcheck**	**Genes identificados**
Donor.bin.47	5,43	79,53	8,56	95	52	5.506
Donor.bin.47 remontado	5,81	84,39	4,26	58	4	5.622
Diferença (%)	**+7,05**	**+4,86**	**-4,30**	**-38,95**	**-92,31**	**+2,11**

B.2 Comparações do artigo GI

B.2.1 *Clostridium baratii*

Tabela B.9: *Artigo GI: Clostridium baratii - Dados dos genomas comparados*

	Tipo	**Tamanho (Mpb)**	**Completude (%)**	**Contaminação (%)**	**Qtde de genes**
dasN1_003_000G1_maxbin2.maxbin.009	MAG	3,21	99,19	1,61	3.004
CDC51267 (GCF_001991075.2)	Referência	3,18	99,19	2,42	3.060
sullivan (GCF_000789395.1)	Referência	3,34	99,19	3,02	3.171

Tabela B.10: *Artigo GI: Clostridium baratii - MAG versus referência*

Código	**ANI** (%)	**RGIs detectadas**		**RGIs detectadas pelo MAGcheck**	
		Qtde	**Tamanho total (pb)**	**Qtde**	**Tamanho total (pb)**
dasN1_003_000G1_maxbin2.maxbin.009	-	16	216.513	-	-
CDC51267	96,01	19	294.651	0	0
sullivan	98,98	15	379.196	0	0

B.2.2 *Enterococcus faecalis*

Tabela B.11: *Artigo GI: Enterococcus faecalis - Dados dos genomas comparados*

	Tipo	**Tamanho (Mpb)**	**Completude (%)**	**Contaminação (%)**	**Qtde de genes**
dasN1_003_019G1	MAG	2,88	99,63	0,00	2.775
EnGen0107 (GCF_000391485.2)	Referência	3,27	99,53	0,19	3.176

Tabela B.12: *Artigo GI: Enterococcus faecalis - MAG versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
dasN1_003_019G1	-	15	213.077	-	-
EnGen0107	99,04	20	566.124	0	0

B.2.3 *Enterococcus faecium*

Tabela B.13: *Artigo GI: Enterococcus faecium - Dados dos genomas comparados*

	Tipo	Tamanho (Mpb)	Completude (%)	Contaminação (%)	Qtde de genes
dasN1_003_000G1_concoct_24	MAG	2,36	98,19	0,47	2.381
dasN1_003_000G1_concoct_24 remontado	MAG	2,61	97,59	1,90	2.630
ISMMS_VRE_1 (GCF_001720945.1)	Referência	3,26	99,63	0,75	3.184

Tabela B.14: *Artigo GI: Enterococcus faecium - MAG versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
dasN1_003_000G1_concoct_24	-	1	5.089	-	-
ISMMS_VRE_1	99,56	47	778.775	27	403.759

Tabela B.15: *Artigo GI: Enterococcus faecium - MAG remontado versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
dasN1_003_000G1_concoct_24 remontado	-	1	6.284	-	-
ISMMS_VRE_1	99,39	32	357.200	10	76.681

Tabela B.16: *Artigo GI: Enterococcus faecium: MAG versus MAG remontado*

	Tamanho (Mpb)	Completude (%)	Contaminação (%)	RGIs negativas	RGIs MAGcheck	Genes identificados
dasN1_003_000G1_concoct_24	2,36	98,19	0,47	47	27	2.381
dasN1_003_000G1_concoct_24 remontado	2,61	97,59	1,90	32	10	2.630
Diferença (%)	**+10,51**	**-0,60**	**+1,43**	**-31,91**	**-62,96**	**+10,46**

B.3 Comparações do artigo HS

B.3.1 *Ardenticatena maritima*

Tabela B.17: *Artigo HS: Ardenticatena maritima - Dados dos genomas comparados*

	Tipo	Tamanho (Mpb)	Completude (%)	Contaminação (%)	Qtde de genes
J129	MAG	2,61	70,22	3,31	2.675
J129 remontado	MAG	2,99	86,30	3,64	3.010
110S.ref (GCF_001306175.1)	Referência	3,56	98,18	4,55	3.032

Tabela B.18: *Artigo HS: Ardenticatena maritima - MAG versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
J129	-	11	95.573	-	-
110S.ref	96,58	69	670.140	64	584.878

Tabela B.19: *Artigo HS: Ardenticatena maritima - MAG remontado versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
J129 remontado	-	11	101.037	-	-
110S.ref	96,57	13	108.947	6	35.951

Tabela B.20: *Artigo HS: Ardenticatena maritima: MAG versus MAG remontado*

	Tamanho (Mpb)	Completude (%)	Contaminação (%)	RGIs negativas	RGIs MAGcheck	Genes identificados
J129	2,61	70,22	3,31	69	64	2.675
J129 remontado	2,99	86,30	3,64	13	6	3.010
Diferença (%)	**+14,78**	**+16,08**	**+0,33**	**-81,16**	**-90,63**	**+12,52**

B.3.2 *Chloroflexus aurantiacus*

Tabela B.21: *Artigo HS: Chloroflexus aurantiacus - Dados dos genomas comparados*

	Tipo	**Tamanho (Mpb)**	**Completude (%)**	**Contaminação (%)**	**Qtde de genes**
J043	MAG	4,60	92,14	0,94	3.924
J043 remontado	MAG	4,91	95,91	0,00	4.127
Y_400_fl (GCF_000022185.1)	Referência	5,27	99,69	0,00	4.338

Tabela B.22: *Artigo HS: Chloroflexus aurantiacus - MAG versus referência*

Código	**ANI** (%)	**RGIs detectadas**		**RGIs detectadas pelo MAGcheck**	
		Qtde	**Tamanho total (pb)**	**Qtde**	**Tamanho total (pb)**
J043	-	4	29.291	-	-
Y_400_fl	99,77	50	526.461	46	474.417

Tabela B.23: *Artigo HS: Chloroflexus aurantiacus - MAG remontado versus referência*

Código	**ANI** (%)	**RGIs detectadas**		**RGIs detectadas pelo MAGcheck**	
		Qtde	**Tamanho total (pb)**	**Qtde**	**Tamanho total (pb)**
J043 remontado	-	5	38.351	-	-
Y_400_fl	99,74	8	93.913	4	47.929

Tabela B.24: *Artigo HS: Chloroflexus aurantiacus: MAG versus MAG remontado*

	Tamanho (Mpb)	**Completude (%)**	**Contaminação (%)**	**RGIs negativas**	**RGIs MAGcheck**	**Genes identificados**
J043	4,60	92,14	0,94	50	46	3.924
J043 remontado	4,91	95,91	0,00	8	4	4.127
Diferença (%)	**+6,82**	**+3,77**	**-0,94**	**-84,00**	**-91,30**	**+5,17**

B.3.3 *Thermosynechococcus sp*

Tabela B.25: *Artigo HS: Thermosynechococcus sp - Dados dos genomas comparados*

	Tipo	**Tamanho (Mpb)**	**Completude (%)**	**Contaminação (%)**	**Qtde de genes**
J003	MAG	2,31	97,96	0,94	2.387
J003 remontado	MAG	2,41	98,43	0,00	2.347
NK55 (GCF_000505665.1)	Referência	2,52	99,76	0,00	2.467

Tabela B.26: *Artigo HS: Thermosynechococcus sp - MAG versus referência*

Código	**ANI** (%)	**RGIs detectadas**		**RGIs detectadas pelo MAGcheck**	
		Qtde	**Tamanho total (pb)**	**Qtde**	**Tamanho total (pb)**
J003	-	0	0	-	-
NK55	99,77	12	109.810	12	109.810

Tabela B.27: *Artigo HS: Thermosynechococcus sp - MAG remontado versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
J003 remontado	-	0	0	-	-
NK55	99,74	0	0	0	0

Tabela B.28: *Artigo HS: Thermosynechococcus sp: MAG versus MAG remontado*

	Tamanho (Mpb)	Completude (%)	Contaminação (%)	RGIs negativas	RGIs MAGcheck	Genes identificados
J003	2,31	97,96	0,94	12	12	2.387
J003 remontado	2,41	98,43	0,00	0	0	2.347
Diferença (%)	**+4,25**	**+0,47**	**-0,94**	**-100,00**	**-100,00**	**-1,68**

B.4 Comparações do artigo PP

B.4.1 *Marivivens sp*

Tabela B.29: *Artigo PP: Marivivens sp - Dados dos genomas comparados*

	Tipo	Tamanho (Mpb)	Completude (%)	Contaminação (%)	Qtde de genes
HF_Dia39	MAG	2,49	95,36	0,95	2.593
HF_Dia39 remontado (com MEGAHIT)	MAG	2,77	96,85	2,10	2.856
JLT3646 (GCF_001908835.1)	Referência	3,15	99,39	0,40	3.133

Tabela B.30: *Artigo PP: Marivivens sp - MAG versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
HF_Dia39	-	3	42.889	-	-
JLT3646	97,96	19	696.613	6	71.053

Tabela B.31: *Artigo PP: Marivivens sp - MAG remontado versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
HF_Dia39 remontado	-	3	42.912	-	-
JLT3646	97,82	18	368.440	1	7.553

Tabela B.32: *Artigo PP: Marivivens sp: MAG versus MAG remontado*

	Tamanho (Mpb)	Completude (%)	Contaminação (%)	RGIs negativas	RGIs MAGcheck	Genes identificados
HF_Dia39	2,49	95,36	0,95	19	6	2.593
HF_Dia39 remontado	2,77	96,85	2,10	18	1	2.856
Diferença	**+11,16**	**+1,49**	**+1,15**	**-5,26**	**-83,33**	**+10,14**

B.4.2 *Ruegeria pomeroyi*

Tabela B.33: *Artigo PP: Ruegeria pomeroyi - Dados dos genomas comparados*

	Tipo	**Tamanho (Mpb)**	**Completude (%)**	**Contaminação (%)**	**Qtde de genes**
HF_Dia28	MAG	4,51	97,56	2,18	4.518
HF_Dia28 remontado	MAG	4,55	98,22	1,28	4.434
DSS_3 (GCF_000011965.2)	Referência	4,60	99,36	0,00	4.393

Tabela B.34: *Artigo PP: Ruegeria pomeroyi - MAG versus referência*

Código	**ANI** (%)	**RGIs detectadas**		**RGIs detectadas pelo MAGcheck**	
		Qtde	**Tamanho total (pb)**	**Qtde**	**Tamanho total (pb)**
HF_Dia28	-	1	6.765	-	-
DSS_3	99,97	7	46.167	7	46.167

Tabela B.35: *Artigo PP: Ruegeria pomeroyi - MAG remontado versus referência*

Código	**ANI** (%)	**RGIs detectadas**		**RGIs detectadas pelo MAGcheck**	
		Qtde	**Tamanho total (pb)**	**Qtde**	**Tamanho total (pb)**
HF_Dia28 remontado	-	1	5.979	-	-
DSS_3	100,00	3	18.022	3	18.022

Tabela B.36: *Artigo PP: Ruegeria pomeroyi: MAG versus MAG remontado*

	Tamanho (Mpb)	Completude (%)	Contaminação (%)	RGIs negativas	RGIs MAGcheck	Genes identificados
HF_Dia28	4,51	97,56	2,18	7	7	4.518
HF_Dia28 remontado	4,55	98,22	1,28	3	3	4.434
Diferença (%)	**+0,77**	**+0,66**	**-0,90**	**-57,14**	**-57,14**	**-1,86**

B.5 Comparações do artigo SG

B.5.1 *Enterococcus faecalis*

Tabela B.37: *Artigo SG: Enterococcus faecalis - Dados dos genomas comparados*

	Tipo	Tamanho (Mpb)	Completude (%)	Contaminação (%)	Qtde de genes
HP_003.bin.12	MAG	2,91	99,63	0,19	2.776
HP_003.bin.12 remontado	MAG	2,94	99,63	0,14	2.795
EnGen0107	Referência	3,27	99,53	0,19	3.176

Tabela B.38: *Artigo SG: Enterococcus faecalis - MAG versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
HP_003.bin.12	-	21	211.604	-	-
EnGen0107	99,05	27	583.812	4	22.577

Tabela B.39: *Artigo SG: Enterococcus faecalis - MAG remontado versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
HP_003.bin.12 remontado	-	18	189.710	-	-
EnGen0107	99,04	26	527.772	1	12.334

Tabela B.40: *Artigo SG: Enterococcus faecalis: MAG versus MAG remontado*

	Tamanho (Mpb)	Completude (%)	Contaminação (%)	RGIs negativas	RGIs MAGcheck	Genes identificados
HP_003.bin.12	2,91	99,63	0,19	27	4	2.776
HP_003.bin.12 remontado	2,94	99,63	0,14	26	1	2.795
Diferença (%)	**+1,14**	**0,00**	**-0,05**	**-3,70**	**-75,00**	**+0,68**

B.5.2 *Enterococcus faecium*

Tabela B.41: *Artigo SG: Enterococcus faecium - Dados dos genomas comparados*

	Tipo	Tamanho (Mpb)	Completude (%)	Contaminação (%)	Qtde de genes
HP_003.bin.36	MAG	2,54	99,63	0,94	2.467
HP_003.bin.36 remontado	MAG	2,62	99,63	0,13	2.550
ISMMS_VRE_1	Referência	3,26	99,63	0,75	3.184

Tabela B.42: *Artigo SG: Enterococcus faecium - MAG versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
HP_003.bin.36	-	15	143.116	-	-
ISMMS_VRE_1	99,56	37	690.107	5	58.965

Tabela B.43: *Artigo SG: Enterococcus faecium - MAG remontado versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
HP_003.bin.36 remontado	-	15	132.020	-	-
ISMMS_VRE_1	99,39	38	538.403	0	0

Tabela B.44: *Artigo SG: Enterococcus faecium: MAG versus MAG remontado*

	Tamanho (Mpb)	Completude (%)	Contaminação (%)	RGIs negativas	RGIs MAGcheck	Genes identificados
HP_003.bin.36	2,54	99,63	0,94	37	5	2.467
HP_003.bin.36 remontado	2,62	99,63	0,13	38	0	2.550
Diferença (%)	**+3,10**	**0,00**	**-0,81**	**+2,70**	**-100,00**	**+3,36**

B.6 Comparações do Metazoo - Bugios

B.6.1 *Prevotella sp*

Tabela B.45: *Metazoo - Bugios: Prevotella sp - Dados dos genomas comparados*

	Tipo	**Tamanho (Mpb)**	**Completude (%)**	**Contaminação (%)**	**Qtde de genes**
bin.8	MAG	3,26	99,26	1,11	2.736
bin.8 remontado	MAG	3,38	99,26	1,16	2.900
P5_50 (GCF_002251295.1)	Referência	2,98	99,26	0,00	2.516

Tabela B.46: *Metazoo - Bugios: Prevotella sp - MAG versus referência*

Código	**ANI** (%)	**RGIs detectadas**		**RGIs detectadas pelo MAGcheck**	
		Qtde	**Tamanho total (pb)**	**Qtde**	**Tamanho total (pb)**
bin.8	-	60	594.557	-	-
P5_50	97,99	32	396.439	1	9.261

Tabela B.47: *Metazoo - Bugios: Prevotella sp - MAG remontado versus referência*

Código	**ANI** (%)	**RGIs detectadas**		**RGIs detectadas pelo MAGcheck**	
		Qtde	**Tamanho total (pb)**	**Qtde**	**Tamanho total (pb)**
bin.8 remontado	-	59	613.584	-	-
P5_50	97,99	29	342.641	0	0

Tabela B.48: *Metazoo - Bugios: Prevotella sp: MAG versus MAG remontado*

	Tamanho (Mpb)	Completude (%)	Contaminação (%)	RGIs negativas	RGIs MAGcheck	Genes identificados
bin.8	3,26	99,26	1,11	32	1	2.736
bin.8 remontado	3,38	99,26	1,16	29	0	2.900
Diferença (%)	**+3,89**	**0,00**	**+0,05**	**-9,38**	**-100,00**	**+5,99**

B.6.2 *Treponema berlinense*

Tabela B.49: *Metazoo - Bugios: Treponema berlinense - Dados dos genomas comparados*

	Tipo	Tamanho (Mpb)	Completude (%)	Contaminação (%)	Qtde de genes
bin.7	MAG	2,34	87,57	0,93	2.156
bin.7 remontado	MAG	2,46	85,78	5,48	2.449
ATCC_BAA_909 (GCF_900167025.1)	Referência	2,52	97,55	0,00	2.175

Tabela B.50: *Metazoo - Bugios: Treponema berlinense - MAG versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
bin.7	-	15	222.923	-	-
ATCC_BAA_909	98,47	27	364.754	2	20.479

Tabela B.51: *Metazoo - Bugios: Treponema berlinense - MAG remontado versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
bin.7 remontado	-	15	237.691	-	-
ATCC_BAA_909	98,47	22	291.218	0	0

Tabela B.52: *Metazoo - Bugios: Treponema berlinense: MAG versus MAG remontado*

	Tamanho (Mpb)	Completude (%)	Contaminação (%)	RGIs negativas	RGIs MAGcheck	Genes identificados
bin.7	2,34	87,57	0,93	27	2	2.156
bin.7 remontado	2,46	85,78	5,48	22	0	2.449
Diferença (%)	**+5,18**	**-1,79**	**+4,55**	**-18,52**	**-100,00**	**+13,59**

B.6.3 *Treponema succinifaciens*

Tabela B.53: *Metazoo - Bugios: Treponema succinifaciens - Dados dos genomas comparados*

	Tipo	Tamanho (Mpb)	Completude (%)	Contaminação (%)	Qtde de genes
bin.22	MAG	2,18	90,05	1,93	2.435
bin.22 remontado	MAG	2,28	90,05	1,28	2.519
DSM_2489 (GCF_000195275.1)	Referência	2,90	100,00	0,00	2.753

Tabela B.54: *Metazoo - Bugios: Treponema succinifaciens - MAG versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
bin.22	-	16	138.822	-	-
DSM_2489	98,76	36	829.801	6	61.822

Tabela B.55: *Metazoo - Bugios: Treponema succinifaciens - MAG remontado versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
bin.22 remontado	-	16	137.302	-	-
DSM_2489	98,72	25	647.719	0	0

Tabela B.56: *Metazoo - Bugios: Treponema succinifaciens: MAG versus MAG remontado*

	Tamanho (Mpb)	Completude (%)	Contaminação (%)	RGIs negativas	RGIs MAGcheck	Genes identificados
bin.22	2,18	90,05	1,93	36	6	2.435
bin.22 remontado	2,28	90,05	1,28	25	0	2.519
Diferença (%)	**+4,20**	**0,00**	**-0,65**	**-30,56**	**-100,00**	**+3,45**

B.7 Comparações do Metazoo - Composto maduro

B.7.1 *Mycolicibacterium thermoresistibile*

Tabela B.57: *Metazoo - Composto maduro: Mycolicibacterium thermoresistibile - Dados dos genomas comparados*

	Tipo	**Tamanho (Mpb)**	**Completude (%)**	**Contaminação (%)**	**Qtde de genes**
ZCMRG14	MAG	3,73	80,59	3,07	4.151
ZCMRG14 remontado	MAG	4,08	87,59	2,58	4.466
NCTC10409 (GCF_900187065.1)	Referência	4,95	99,32	0,00	4.672

Tabela B.58: *Metazoo - Composto maduro: Mycolicibacterium thermoresistibile - MAG versus referência*

Código	**ANI** (%)	**RGIs detectadas**		**RGIs detectadas pelo MAGcheck**	
		Qtde	**Tamanho total (pb)**	**Qtde**	**Tamanho total (pb)**
ZCMRG14	-	3	24.844	-	-
NCTC10409	99,75	64	939.991	46	417.654

Tabela B.59: *Metazoo - Composto maduro: Mycolicibacterium thermoresistibile - MAG remontado versus referência*

Código	**ANI** (%)	**RGIs detectadas**		**RGIs detectadas pelo MAGcheck**	
		Qtde	**Tamanho total (pb)**	**Qtde**	**Tamanho total (pb)**
ZCMRG14 remontado	-	3	24.881	-	-
NCTC10409	99,74	20	466.457	2	18.617

Tabela B.60: *Metazoo - Composto maduro: Mycolicibacterium thermoresistibile: MAG versus MAG remontado*

	Tamanho (Mpb)	Completude (%)	Contaminação (%)	RGIs negativas	RGIs MAGcheck	Genes identificados
ZCMRG14	3,73	80,59	3,07	64	46	4.151
ZCMRG14 remontado	4,08	87,59	2,58	20	2	4.466
Diferença (%)	**+9,47**	**+7,00**	**-0,49**	**-68,75**	**-95,65**	**+7,59**

B.7.2 *Novibacillus thermophilus*

Tabela B.61: *Metazoo - Composto maduro: Novibacillus thermophilus - Dados dos genomas comparados*

	Tipo	Tamanho (Mpb)	Completude (%)	Contaminação (%)	Qtde de genes
bin.6	MAG	1,54	60,66	2,69	1.881
bin.6 remontado	MAG	2,37	73,40	3,46	3.022
SG_1 (GCF_002005165.1)	Referência	3,63	98,08	0,00	3.683

Tabela B.62: *Metazoo - Composto maduro: Novibacillus thermophilus - MAG versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
bin.6	-	1	7.982	-	-
SG_1 (GCF_002005165.1)	99,14	109	1.629.372	97	1.118.528

Tabela B.63: *Metazoo - Composto maduro: Novibacillus thermophilus - MAG remontado versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
bin.6 remontado	-	1	7.978	-	-
SG_1 (GCF_002005165.1)	99,17	25	270.054	1	6.104

Tabela B.64: *Metazoo - Composto maduro: Novibacillus thermophilus: MAG versus MAG remontado*

	Tamanho (Mpb)	Completude (%)	Contaminação (%)	RGIs negativas	RGIs MAGcheck	Genes identificados
bin.6	1,54	60,66	2,69	109	97	1.881
bin.6 remontado	2,37	73,40	3,46	25	1	3.022
Diferença (%)	**+53,44**	**+12,74**	**+0,77**	**-77,06**	**-98,97**	**+60,66**

B.7.3 *Planifilum fulgidum*

Tabela B.65: *Metazoo - Composto maduro: Planifilum fulgidum - Dados dos genomas comparados*

	Tipo	Tamanho (Mpb)	Completude (%)	Contaminação (%)	Qtde de genes
bin.9	MAG	2,05	62,72	5,67	2.789
bin.9 remontado	MAG	2,54	76,50	4,45	3.291
DSM_44945 (GCF_900113175.1)	Referência	3,38	97,69	0,51	3.262

Tabela B.66: *Metazoo - Composto maduro: Planifilum fulgidum - MAG versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
bin9	-	0	0	-	-
DSM_44945	96,58	112	1.017.830	102	928.578

Tabela B.67: *Metazoo - Composto maduro: Planifilum fulgidum - MAG remontado versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
bin.9 remontado	-	0	0	-	-
DSM_44945	97,11	21	153.210	7	39.293

Tabela B.68: *Metazoo - Composto maduro: Planifilum fulgidum: MAG versus MAG remontado*

	Tamanho (Mpb)	Completude (%)	Contaminação (%)	RGIs negativas	RGIs MAGcheck	Genes identificados
bin.9	2,05	62,72	5,67	112	102	2.789
bin.9 remontado	2,54	76,50	4,45	21	7	3.291
Diferença (%)	**+24,06**	**+13,78**	**-1,22**	**-81,25**	**-93,14**	**+18,00**

B.7.4 *Thermocrispum municipale*

Tabela B.69: *Metazoo - Composto maduro: Thermocrispum municipale - Dados dos genomas comparados*

	Tipo	**Tamanho (Mpb)**	**Completude (%)**	**Contaminação (%)**	**Qtde de genes**
ZCMRG04	MAG	3,73	80,59	3,07	4.151
ZCMRG04 remontado	MAG	4,08	87,59	2,58	4.466
DSM_44069 (GCF_000427825.1)	Referência	4,95	99,32	0,00	4.672

Tabela B.70: *Metazoo - Composto maduro: Thermocrispum municipale - MAG versus referência*

Código	**ANI** (%)	**RGIs detectadas**		**RGIs detectadas pelo MAGcheck**	
		Qtde	**Tamanho total (pb)**	**Qtde**	**Tamanho total (pb)**
ZCMRG04	-	22	284.560	-	-
DSM_44069	99,53	20	392.384	12	256.450

Tabela B.71: *Metazoo - Composto maduro: Thermocrispum municipale - MAG remontado versus referência*

Código	**ANI** (%)	**RGIs detectadas**		**RGIs detectadas pelo MAGcheck**	
		Qtde	**Tamanho total (pb)**	**Qtde**	**Tamanho total (pb)**
ZCMRG04 remontado	-	21	284.177	-	-
DSM_44069	99,52	10	145.605	2	14.092

Tabela B.72: *Metazoo - Composto maduro: Thermocrispum municipale: MAG versus MAG remontado*

	Tamanho (Mpb)	Completude (%)	Contaminação (%)	RGIs negativas	RGIs MAGcheck	Genes identificados
ZCMRG04	3,73	80,59	3,07	20	12	4.151
ZCMRG04 remontado	4,08	87,59	2,58	10	2	4.466
Diferença (%)	**+9,47**	**+7,00**	**-0,49**	**-50,00**	**-83,33**	**+7,59**

B.8 Comparações do Metazoo - Inóculo

B.8.1 *Sphaerobacter thermophilus*

Tabela B.73: *Metazoo - Inóculo: Sphaerobacter thermophilus - Dados dos genomas comparados*

	Tipo	Tamanho (Mpb)	Completude (%)	Contaminação (%)	Qtde de genes
bin.12	MAG	2,42	68,80	4,52	2.714
bin.12 remontado	MAG	3,03	81,58	6,45	3.450
DSM_20745 (GCF_000024985.1)	Referência	3,99	98,13	0,00	3.523

Tabela B.74: *Metazoo - Inóculo: Sphaerobacter thermophilus - MAG versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
bin.12	-	10	64.421	-	-
DSM_20745	98,88	128	1.180.956	107	931.365

Tabela B.75: *Metazoo - Inóculo: Sphaerobacter thermophilus - MAG remontado versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
bin.12 remontado	-	8	51.585	-	-
DSM_20745	98,85	23	186.159	4	27.344

Tabela B.76: *Metazoo - Inóculo: Sphaerobacter thermophilus: MAG versus MAG remontado*

	Tamanho (Mpb)	Completude (%)	Contaminação (%)	RGIs negativas	RGIs MAGcheck	Genes identificados
bin.12	2,42	68,80	4,52	128	107	2.714
bin.12 remontado	3,03	81,58	6,45	23	4	3.450
Diferença (%)	**+24,91**	**+12,78**	**+1,93**	**-82,03**	**-96,26**	**+27,12**

B.9 Comparações do Metazoo - Lago

B.9.1 *Limnohabitans sp*

Tabela B.77: *Metazoo - Lago: Limnohabitans sp - Dados dos genomas comparados*

	Tipo	Tamanho (Mpb)	Completude (%)	Contaminação (%)	Qtde de genes
ZLKRG25	MAG	2,62	97,27	3,80	2.686
ZLKRG25 remontado	MAG	2,81	97,89	3,22	2.852
103DPR2 (GCF_001412575.1)	Referência	2,95	99,22	1,95	2.806

Tabela B.78: *Metazoo - Lago: Limnohabitans sp - MAG versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
ZLKRG25	-	12	89.237	-	-
103DPR2	97,96	26	442.506	14	211.455

Tabela B.79: *Metazoo - Lago: Limnohabitans sp - MAG remontado versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
ZLKRG25 remontado	-	7	76.505	-	-
103DPR2	98,06	16	199.522	1	5.850

Tabela B.80: *Metazoo - Lago: Limnohabitans sp: MAG versus MAG remontado*

	Tamanho (Mpb)	Completude (%)	Contaminação (%)	RGIs negativas	RGIs MAGcheck	Genes identificados
ZLKRG25	2,62	97,27	3,80	26	14	2.686
ZLKRG25 remontado	2,81	97,89	3,22	16	1	2.852
Diferença (%)	**+7,46**	**+0,62**	**-0,58**	**-38,46**	**-92,86**	**+6,18**

B.10 Comparações do Metazoo - ZC3

B.10.1 *Planifilum fulgidum*

Tabela B.81: *Metazoo - ZC3: Planifilum fulgidum - Dados dos genomas comparados*

	Tipo	**Tamanho (Mpb)**	**Completude (%)**	**Contaminação (%)**	**Qtde de genes**
ZC3RG08	MAG	3,03	93,08	0,00	2.988
ZC3RG08 remontado	MAG	3,26	97,69	0,77	3.196
DSM_44945 (GCF_900113175.1)	Referência	3,38	97,69	0,51	3.262

Tabela B.82: *Metazoo - ZC3: Planifilum fulgidum - MAG versus referência*

Código	**ANI** (%)	**RGIs detectadas**		**RGIs detectadas pelo MAGcheck**	
		Qtde	**Tamanho total (pb)**	**Qtde**	**Tamanho total (pb)**
ZC3RG08	-	5	36.182	-	-
DSM_44945	99,41	19	325.864	13	272.645

Tabela B.83: *Metazoo - ZC3: Planifilum fulgidum - MAG remontado versus referência*

Código	**ANI** (%)	**RGIs detectadas**		**RGIs detectadas pelo MAGcheck**	
		Qtde	**Tamanho total (pb)**	**Qtde**	**Tamanho total (pb)**
ZC3RG08 remontado	-	6	41.875	-	-
DSM_44945	99,39	4	39.295	0	0

Tabela B.84: *Metazoo - ZC3: Planifilum fulgidum: MAG versus MAG remontado*

	Tamanho (Mpb)	Completude (%)	Contaminação (%)	RGIs negativas	RGIs MAGcheck	Genes identificados
ZC3RG08	3,03	93,08	0,00	19	13	2.988
ZC3RG08 remontado	3,26	97,69	0,77	4	0	3.196
Diferença (%)	**+7,64**	**+4,61**	**+0,77**	**-78,95**	**-100,00**	**+6,96**

B.10.2 *Thermobifida fusca*

Tabela B.85: *Metazoo - ZC3: Thermobifida fusca - Dados dos genomas comparados*

	Tipo	Tamanho (Mpb)	Completude (%)	Contaminação (%)	Qtde de genes
ZC3RG05	MAG	3,61	99,45	0,00	3.171
ZC3RG05 remontado	MAG	3,63	99,45	0,00	3.192
UPMC_901 (GCF_015034585.1)	Referência	3,76	98,91	0,00	3.290
YX (GCF_000012405.1)	Referência	3,64	99,45	0,00	3.146

Tabela B.86: *Metazoo - ZC3: Thermobifida fusca - MAG versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
ZC3RG05	-	3	27.111	-	-
UPMC_901	99,76	12	179.142	5	28.885
YX	99,80	8	48.272	6	34.247

Tabela B.87: *Metazoo - ZC3: Thermobifida fusca - MAG remontado versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
ZC3RG05 remontado	-	3	27.123	-	-
UPMC_901	99,78	5	133.596	0	0
YX	99,81	0	0	0	0

Tabela B.88: *Metazoo - ZC3: Thermobifida fusca: MAG versus MAG remontado*

	Tamanho (Mpb)	Completude (%)	Contaminação (%)	RGIs negativas	RGIs MAGcheck	Genes identificados
ZC3RG05	3,61	99,45	0,00	20	11	3.171
ZC3RG05 remontado	3,63	99,45	0,00	5	0	3.192
Diferença (%)	**+0,71**	**0,00**	**0,00**	**-75,00**	**-100,00**	**+0,66**

B.11 Comparações do Metazoo - ZC4

B.11.1 *Caldibacillus debilis*

Tabela B.89: *Metazoo - ZC4: Caldibacillus debilis - Dados dos genomas comparados*

	Tipo	Tamanho (Mpb)	Completude (%)	Contaminação (%)	Qtde de genes
ZC4RG01	MAG	2,80	91,47	0,26	2.587
ZC4RG01 remontado	MAG	3,07	91,59	0,68	2.879
B4135 (GCF_001587535.1)	Referência	3,25	91,69	0,25	3.055
DSM_16016 (GCF_000383875.1)	Referência	3,06	91,69	0,25	2.805
GB1 (GCF_003627895.1)	Referência	3,35	91,58	3,42	3.261

Tabela B.90: *Metazoo - ZC4: Caldibacillus debilis - MAG versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
ZC4RG01	-	0	0	-	-
B4135	98,86	30	366.851	18	186.555
DSM_16016	99,01	26	218.876	18	149.251
GB1	98,75	29	530.855	17	168.081

Tabela B.91: *Metazoo - ZC4: Caldibacillus debilis - MAG remontado versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
ZC4RG01 remontado	-	0	0	-	-
B4135	98,83	17	208.265	4	31.355
DSM_16016	99,00	9	70.775	1	10.801
GB1	98,72	16	340.058	2	19.334

Tabela B.92: *Metazoo - ZC4: Caldibacillus debilis: MAG versus MAG remontado*

	Tamanho (Mpb)	Completude (%)	Contaminação (%)	RGIs negativas	RGIs MAGcheck	Genes identificados
ZC4RG01	2,80	91,47	0,26	85	53	2.587
ZC4RG01 remontado	3,07	91,59	0,68	42	7	2.879
Diferença (%)	**+9,76**	**+0,12**	**+0,42**	**-50,59**	**-86,79**	**+11,29**

B.11.2 *Caldicoprobacter oshimai*

Tabela B.93: *Metazoo - ZC4: Caldicoprobacter oshimai - Dados dos genomas comparados*

	Tipo	**Tamanho (Mpb)**	**Completude (%)**	**Contaminação (%)**	**Qtde de genes**
ZC4RG32	MAG	2,69	92,72	6,72	2.812
ZC4RG32 remontado	MAG	2,73	93,52	6,88	2.673
DSM_21659 (GCF_000526435.1)	Referência	2,73	96,75	1,21	2.525

Tabela B.94: *Metazoo - ZC4: Caldicoprobacter oshimai - MAG versus referência*

Código	**ANI** (%)	**RGIs detectadas**		**RGIs detectadas pelo MAGcheck**	
		Qtde	**Tamanho total (pb)**	**Qtde**	**Tamanho total (pb)**
ZC4RG32	-	20	226.507	-	-
DSM_21659	99,16	40	502.009	26	277.271

Tabela B.95: *Metazoo - ZC4: Caldicoprobacter oshimai - MAG remontado versus referência*

Código	**ANI** (%)	**RGIs detectadas**		**RGIs detectadas pelo MAGcheck**	
		Qtde	**Tamanho total (pb)**	**Qtde**	**Tamanho total (pb)**
ZC4RG32 remontado	-	24	259.850	-	-
DSM_21659	99,15	29	292.955	17	128.897

Tabela B.96: *Metazoo - ZC4: Caldicoprobacter oshimai: MAG versus MAG remontado*

	Tamanho (Mpb)	Completude (%)	Contaminação (%)	RGIs negativas	RGIs MAGcheck	Genes identificados
ZC4RG32	2,69	92,72	6,72	40	26	2.812
ZC4RG32 remontado	2,73	93,52	6,88	29	17	2.673
Diferença (%)	**+1,20**	**+0,80**	**+0,16**	**-27,50**	**-34,62**	**-4,94**

B.11.3 *Clostridium cellulosi*

Tabela B.97: *Metazoo - ZC4: Clostridium cellulosi - Dados dos genomas comparados*

	Tipo	Tamanho (Mpb)	Completude (%)	Contaminação (%)	Qtde de genes
ZC4RG49	MAG	1,59	80,35	2,17	1.798
ZC4RG49 remontado	MAG	1,76	83,71	0,89	1.874
DG5 (GCF_000953215.1)	Referência	2,23	98,32	0,67	2.017

Tabela B.98: *Metazoo - ZC4: Clostridium cellulosi - MAG versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
ZC4RG49	-	0	0	-	-
DG5	99,69	36	371.654	27	269.348

Tabela B.99: *Metazoo - ZC4: Clostridium cellulosi - MAG remontado versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
ZC4RG49 remontado	-	0	0	-	-
DG5	99,66	10	91.787	2	14.357

Tabela B.100: *Metazoo - ZC4: Clostridium cellulosi: MAG versus MAG remontado*

	Tamanho (Mpb)	Completude (%)	Contaminação (%)	RGIs negativas	RGIs MAGcheck	Genes identificados
ZC4RG49	1,59	80,35	2,17	36	27	1.798
ZC4RG49 remontado	1,76	83,71	0,89	10	2	1.874
Diferença (%)	**+10,70**	**+3,36**	**-1,28**	**-72,22**	**-92,59**	**+4,23**

B.11.4 *Mycobacterium hassiacum*

Tabela B.101: *Metazoo - ZC4: Mycobacterium hassiacum - Dados dos genomas comparados*

	Tipo	Tamanho (Mpb)	Completude (%)	Contaminação (%)	Qtde de genes
ZC4RG43	MAG	4,35	85,03	4,70	5.180
ZC4RG43 remontado	MAG	4,44	86,77	1,97	4.865
DSM_44199 (GCF_900603025.1)	Referência	5,27	99,77	0,15	4.928

Tabela B.102: *Metazoo - ZC4: Mycobacterium hassiacum - MAG versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
ZC4RG43	-	3	27.770	-	-
DSM_44199	99,76	37	437.636	28	270.790

Tabela B.103: *Metazoo - ZC4: Mycobacterium hassiacum - MAG remontado versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
ZC4RG43 remontado	-	3	28.930	-	-
DSM_44199	99,81	13	146.922	2	19.331

Tabela B.104: *Metazoo - ZC4: Mycobacterium hassiacum: MAG versus MAG remontado*

	Tamanho (Mpb)	Completude (%)	Contaminação (%)	RGIs negativas	RGIs MAGcheck	Genes identificados
ZC4RG43	4,35	85,03	4,70	37	28	5.180
ZC4RG43 remontado	4,44	86,77	1,97	13	2	4.865
Diferença (%)	**+2,15**	**+1,74**	**-2,73**	**-64,86**	**-92,86**	**-6,08**

B.11.5 *Planifilum fulgidum*

Tabela B.105: *Metazoo - ZC4: Planifilum fulgidum - Dados dos genomas comparados*

	Tipo	**Tamanho (Mpb)**	**Completude (%)**	**Contaminação (%)**	**Qtde de genes**
ZC4RG09	MAG	3,20	97,69	0,51	3.085
ZC4RG09 remontado	MAG	3,34	97,69	0,51	3.206
DSM_44945 (GCF_900113175.1)	Referência	3,38	97,69	0,51	3.262

Tabela B.106: *Metazoo - ZC4: Planifilum fulgidum - MAG versus referência*

Código	**ANI** (%)	**RGIs detectadas**		**RGIs detectadas pelo MAGcheck**	
		Qtde	**Tamanho total (pb)**	**Qtde**	**Tamanho total (pb)**
ZC4RG09	-	4	28.312	-	-
DSM_44945	99,42	16	190.338	14	166.903

Tabela B.107: *Metazoo - ZC4: Planifilum fulgidum - MAG remontado versus referência*

Código	**ANI** (%)	**RGIs detectadas**		**RGIs detectadas pelo MAGcheck**	
		Qtde	**Tamanho total (pb)**	**Qtde**	**Tamanho total (pb)**
ZC4RG09 remontado	-	4	31.703	-	-
DSM_44945	99,41	4	33.996	2	10.561

Tabela B.108: *Metazoo - ZC4: Planifilum fulgidum: MAG versus MAG remontado*

	Tamanho (Mpb)	Completude (%)	Contaminação (%)	RGIs negativas	RGIs MAGcheck	Genes identificados
ZC4RG09	3,20	97,69	0,51	16	14	3.085
ZC4RG09 remontado	3,34	97,69	0,51	4	2	3.206
Diferença (%)	**+4,45**	**0,00**	**0,00**	**-75,00**	**-85,71**	**+3,92**

B.11.6 *Pseudomonas thermotolerans*

Tabela B.109: *Metazoo - ZC4: Pseudomonas thermotolerans - Dados dos genomas comparados*

	Tipo	Tamanho (Mpb)	Completude (%)	Contaminação (%)	Qtde de genes
ZC4RG08	MAG	3,55	94,23	3,52	3.415
ZC4RG08 remontado	MAG	3,71	95,33	4,81	3.600
DSM_14292 (GCF_000364625.1)	Referência	3,75	98,38	1,24	3.541
J53 (GCF_000513835.1)	Referência	3,75	98,38	0,92	3.507

Tabela B.110: *Metazoo - ZC4: Pseudomonas thermotolerans - MAG versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
ZC4RG08	-	8	67.320	-	-
DSM_14292	99,29	24	283.266	14	163.047
J53	99,15	22	243.429	14	137.892

Tabela B.111: *Metazoo - ZC4: Pseudomonas thermotolerans - MAG remontado versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
ZC4RG08 remontado	-	6	48.256	-	-
DSM_14292	99,29	10	91.843	2	10.293
J53	99,17	11	127.701	3	22.164

Tabela B.112: *Metazoo - ZC4: Pseudomonas thermotolerans: MAG versus MAG remontado*

	Tamanho (Mpb)	Completude (%)	Contaminação (%)	RGIs negativas	RGIs MAGcheck	Genes identificados
ZC4RG08	3,55	94,23	3,52	46	28	3.415
ZC4RG08 remontado	3,71	95,33	4,81	21	5	3.600
Diferença (%)	**+4,68**	**+1,10**	**+1,29**	**-54,35**	**-82,14**	**+5,42**

B.11.7 *Rhodothermus marinus*

Tabela B.113: *Metazoo - ZC4: Rhodothermus marinus - Dados dos genomas comparados*

	Tipo	Tamanho (Mpb)	Completude (%)	Contaminação (%)	Qtde de genes
ZC4RG13	MAG	3,10	99,44	0,00	2.778
ZC4RG13 remontado	MAG	3,27	100,00	0,00	2.901
AA2_13 (GCA_009936255.1)	Referência	3,44	100,00	0,00	2.997
AA3_38 (GCA_009936275.1)	Referência	3,43	100,00	0,00	2.993
DSM_4252 (GCF_000024845.1)	Referência	3,39	100,00	0,00	2.922
SG0_5JP17_172 (GCF_000224745.1)	Referência	3,33	100,00	0,00	2.946

Tabela B.114: *Metazoo - ZC4: Rhodothermus marinus - MAG versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
ZC4RG13	-	2	12.499	-	-
AA2_13	94,38	31	431.683	14	152.068
AA3_38	94,30	28	407.225	11	124.391
DSM_4252	94,81	29	380.012	7	55.683
SG0_5JP17_172	98,82	18	246.131	10	102.169

Tabela B.115: *Metazoo - ZC4: Rhodothermus marinus - MAG remontado versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
ZC4RG13 remontado	-	2	12.669	-	-
AA2_13	94,43	20	309.885	2	20.404
AA3_38	94,30	20	298.323	3	26.384
DSM_4252	94,77	23	283.862	0	0
SG0_5JP17_172	98,82	9	144.598	1	5.417

Tabela B.116: *Metazoo - ZC4: Rhodothermus marinus: MAG versus MAG remontado*

	Tamanho (Mpb)	Completude (%)	Contaminação (%)	RGIs negativas	RGIs MAGcheck	Genes identificados
ZC4RG13	3,10	99,44	0,00	106	42	2.778
ZC4RG13 remontado	3,27	100,00	0,00	72	6	2.901
Diferença (%)	**+5,43**	**+0,56**	**0,00**	**-32,08**	**-85,71**	**+4,43**

B.11.8 *Sphaerobacter thermophilus*

Tabela B.117: *Metazoo - ZC4: Sphaerobacter thermophilus - Dados dos genomas comparados*

	Tipo	**Tamanho (Mpb)**	**Completude (%)**	**Contaminação (%)**	**Qtde de genes**
ZC4RG26	MAG	3,92	95,33	0,93	3.755
ZC4RG26 remontado	MAG	3,99	94,39	0,00	3.672
DSM_20745 (GCF_000024985.1)	Referência	3,99	98,13	0,00	3.523

Tabela B.118: *Metazoo - ZC4: Sphaerobacter thermophilus - MAG versus referência*

Código	**ANI** (%)	**RGIs detectadas**		**RGIs detectadas pelo MAGcheck**	
		Qtde	**Tamanho total (pb)**	**Qtde**	**Tamanho total (pb)**
ZC4RG26	-	13	106.057	-	-
DSM_20745	98,97	24	193.294	19	152.219

Tabela B.119: *Metazoo - ZC4: Sphaerobacter thermophilus - MAG remontado versus referência*

Código	**ANI** (%)	**RGIs detectadas**		**RGIs detectadas pelo MAGcheck**	
		Qtde	**Tamanho total (pb)**	**Qtde**	**Tamanho total (pb)**
ZC4RG26 remontado	-	12	112.075	-	-
DSM_20745	98,95	6	48.795	1	6.840

Tabela B.120: *Metazoo - ZC4: Sphaerobacter thermophilus: MAG versus MAG remontado*

	Tamanho (Mpb)	Completude (%)	Contaminação (%)	RGIs negativas	RGIs MAGcheck	Genes identificados
ZC4RG26	3,92	95,33	0,93	24	19	3.755
ZC4RG26 remontado	3,99	94,39	0,00	6	1	3.672
Diferença (%)	**+1,72**	**-0,94**	**-0,93**	**-75,00**	**-94,74**	**-2,21**

B.11.9 *Thermobifida fusca*

Tabela B.121: *Metazoo - ZC4: Thermobifida fusca - Dados dos genomas comparados*

	Tipo	Tamanho (Mpb)	Completude (%)	Contaminação (%)	Qtde de genes
ZC4RG21	MAG	3,61	99,45	0,55	3.305
ZC4RG21 remontado	MAG	3,63	99,45	0,11	3.210
UPMC_901 (GCF_015034585.1)	Referência	3,76	98,91	0,00	3.290
YX (GCF_000012405.1)	Referência	3,64	99,45	0,00	3.146

Tabela B.122: *Metazoo - ZC4: Thermobifida fusca - MAG versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
ZC4RG21	-	3	25.572	-	-
UPMC_901	98,81	11	172.435	7	56.771
YX	99,86	8	48.255	7	41.226

Tabela B.123: *Metazoo - ZC4: Thermobifida fusca - MAG remontado versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
ZC4RG21 remontado	-	3	26.955	-	-
UPMC_901	99,80	5	113.315	1	5.026
YX	99,86	0	0	0	0

Tabela B.124: *Metazoo - ZC4: Thermobifida fusca: MAG versus MAG remontado*

	Tamanho (Mpb)	Completude (%)	Contaminação (%)	RGIs negativas	RGIs MAGcheck	Genes identificados
ZC4RG21	3,61	99,45	0,55	19	14	3.305
ZC4RG21 remontado	3,63	99,45	0,11	5	1	3.210
Diferença (%)	**+0,55**	**0,00**	**-0,44**	**-73,68**	**-92,86**	**-2,87**

B.11.10 *Thermobispora bispora*

Tabela B.125: *Metazoo - ZC4: Thermobispora bispora - Dados dos genomas comparados*

	Tipo	Tamanho (Mpb)	Completude (%)	Contaminação (%)	Qtde de genes
ZC4RG04	MAG	4,05	97,53	0,79	3.505
ZC4RG04 remontado	MAG	4,21	98,94	0,53	3.632
DSM_43833 (GCF_000092645.1)	Referência	4,19	98,94	0,26	3.588

Tabela B.126: *Metazoo - ZC4: Thermobispora bispora - MAG versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
ZC4RG04	-	3	27.008	-	-
DSM_43833	99,75	10	155.491	10	155.491

Tabela B.127: *Metazoo - ZC4: Thermobispora bispora - MAG remontado versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
ZC4RG04 remontado	-	3	27.175	-	-
DSM_43833	99,74	0	0	0	0

Tabela B.128: *Metazoo - ZC4: Thermobispora bispora: MAG versus MAG remontado*

	Tamanho (Mpb)	Completude (%)	Contaminação (%)	RGIs negativas	RGIs MAGcheck	Genes identificados
ZC4RG04	4,05	97,53	0,79	10	10	3.505
ZC4RG04 remontado	4,21	98,94	0,53	0	0	3.632
Diferença (%)	**+3,80**	**+1,41**	**-0,26**	**-100,00**	**-100,00**	**+3,62**

B.11.11 *Thermocrispum agreste*

Tabela B.129: *Metazoo - ZC4: Thermocrispum agreste - Dados dos genomas comparados*

	Tipo	**Tamanho (Mpb)**	**Completude (%)**	**Contaminação (%)**	**Qtde de genes**
ZC4RG45	MAG	3,82	83,24	3,51	4.296
ZC4RG45 remontado	MAG	3,84	83,69	2,71	3.990
DSM_44070 (GCF_000427905.1)	Referência	4,20	100,00	0,25	3.786

Tabela B.130: *Metazoo - ZC4: Thermocrispum agreste - MAG versus referência*

Código	**ANI** (%)	**RGIs detectadas**		**RGIs detectadas pelo MAGcheck**	
		Qtde	**Tamanho total (pb)**	**Qtde**	**Tamanho total (pb)**
ZC4RG45	-	9	66.484	-	-
DSM_44070	99,72	14	126.014	12	78.617

Tabela B.131: *Metazoo - ZC4: Thermocrispum agreste - MAG remontado versus referência*

Código	**ANI** (%)	**RGIs detectadas**		**RGIs detectadas pelo MAGcheck**	
		Qtde	**Tamanho total (pb)**	**Qtde**	**Tamanho total (pb)**
ZC4RG45 remontado	-	8	65.066	-	-
DSM_44070	99,76	2	23.499	1	5.070

Tabela B.132: *Metazoo - ZC4: Thermocrispum agreste: MAG versus MAG remontado*

	Tamanho (Mpb)	Completude (%)	Contaminação (%)	RGIs negativas	RGIs MAGcheck	Genes identificados
ZC4RG45	3,82	83,24	3,51	14	12	4.296
ZC4RG45 remontado	3,84	83,69	2,71	2	1	3.990
Diferença (%)	**+0,57**	**+0,45**	**-0,80**	**-85,71**	**-91,67**	**-7,12**

B.12 Comparações do SRA

B.12.1 *Desulfurellaceae bacterium*

Tabela B.133: *SRA: Desulfurellaceae bacterium - Dados dos genomas comparados*

	Tipo	Tamanho (Mpb)	Completude (%)	Contaminação (%)	Qtde de genes
bin.39	MAG	2,77	71,30	4,08	3.394
bin.39 remontado	MAG	3,93	77,18	6,61	4.862
bin.18.ref (GCA_002238415.1)	Referência	5,24	88,71	3,36	5.330

Tabela B.134: *SRA: Desulfurellaceae bacterium - MAG versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
bin.39	-	6	37.806	-	-
bin.18.ref	95,67	174	2.017.769	108	1.140.253

Tabela B.135: *SRA: Desulfurellaceae bacterium - MAG remontado versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
bin.39 remontado	-	13	90.816	-	-
bin.18.ref	94,94	110	996.826	37	267.372

Tabela B.136: *SRA: Desulfurellaceae bacterium: MAG versus MAG remontado*

	Tamanho (Mpb)	Completude (%)	Contaminação (%)	RGIs negativas	RGIs MAGcheck	Genes identificados
bin.39	2,77	71,30	4,08	174	108	3.394
bin.39 remontado	3,93	77,18	6,61	110	37	4.862
Diferença (%)	**+41,89**	**+5,88**	**+2,53**	**-36,78**	**-65,74**	**+43,25**

B.12.2 *Truepera sp*

Tabela B.137: *SRA: Truepera sp - Dados dos genomas comparados*

	Tipo	Tamanho (Mpb)	Completude (%)	Contaminação (%)	Qtde de genes
bin.112	MAG	1,73	64,36	3,25	1.837
bin.112 remontado	MAG	2,29	73,89	3,67	2.390
bin.119.ref (GCA_002239005.1)	Referência	3,21	87,50	1,06	3.024

Tabela B.138: *SRA: Truepera sp - MAG versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
bin.112	-	5	34.009	-	-
bin.119.ref	95,95	100	1.220.936	52	471.875

Tabela B.139: *SRA: Truepera sp - MAG remontado versus referência*

Código	ANI (%)	RGIs detectadas		RGIs detectadas pelo MAGcheck	
		Qtde	Tamanho total (pb)	Qtde	Tamanho total (pb)
bin.112 remontado	-	4	24.860	-	-
bin.119.ref	95,88	34	344.036	2	14.074

Tabela B.140: *SRA: Truepera sp: MAG versus MAG remontado*

	Tamanho (Mpb)	Completude (%)	Contaminação (%)	RGIs negativas	RGIs MAGcheck	Genes identificados
bin.112	1,73	64,36	3,25	100	52	1.837
bin.112 remontado	2,29	73,89	3,67	34	2	2.390
Diferença (%)	**+32,54**	**+9,53**	**+0,42**	**-66,00**	**-96,15**	**+30,10**

Referências Bibliográficas

[1] O. V. Goloshchapov, E. I. Olekhnovich, S. V. Sidorenko, I. S. Moiseev, M. A. Kucher, D. E. Fedorov, A. V. Pavlenko, A. I. Manolov, V. V. Gostev, V. A. Veselovsky, K. M. Klimina, E. S. Kostryukova, E. A. Bakin, A. N. Shvetcov, E. D. Gumbatova, R. V. Klementeva, A. A. Shcherbakov, M. V. Gorchakova, J. J. Egozcue, V. Pawlowsky-Glahn, M. A. Suvorova, A. B. Chukhlovin, V. M. Govorun, E. N. Ilina, and B. V. Afanasyev, "Long-term impact of fecal transplantation in healthy volunteers," *BMC Microbiology*, vol. 19, no. 1, dec 2019. xix, 33

[2] M. R. Olm, N. Bhattacharya, A. Crits-Christoph, B. A. Firek, R. Baker, Y. S. Song, M. J. Morowitz, and J. F. Banfield, "Necrotizing enterocolitis is preceded by increased gut bacterial replication, klebsiella, and fimbriae-encoding bacteria," *Science Advances*, vol. 5, no. 12, p. eaax5727, dec 2019. xix, 33

[3] L. M. Ward, A. Idei, M. Nakagawa, Y. Ueno, W. W. Fischer, and S. E. McGlynn, "Geochemical and metagenomic characterization of jinata onsen, a proterozoic-analog hot spring, reveals novel microbial diversity including iron-tolerant phototrophs and thermophilic lithotrophs," *Microbes and Environments*, vol. 34, no. 3, pp. 278–292, 2019. xix, 33

[4] H. Fu, C. B. Smith, S. Sharma, and M. A. Moran, "Genome sequences and metagenome-assembled genome sequences of microbial communities enriched on phytoplankton exometabolites," *Microbiology Resource Announcements*, vol. 9, no. 30, jul 2020. xix, 33

[5] E. I. Olekhnovich, A. I. Manolov, A. E. Samoilov, N. A. Prianichnikov, M. V. Malakhova, A. V. Tyakht, A. V. Pavlenko, V. V. Babenko, A. K. Larin, B. A. Kovarsky, E. V. Starikova, O. E. Glushchenko, D. D. Safina, M. I. Markelova, E. A. Boulygina, D. R. Khusnutdinova, S. Y. Malanin, S. R. Abdulkhakov, R. A. Abdulkhakov, T. V. Grigoryeva, E. S. Kostryukova, V. M. Govorun, and E. N. Ilina, "Shifts in the human gut microbiota structure caused by quadruple helicobacter pylori eradication therapy," *Frontiers in Microbiology*, vol. 10, aug 2019. xix, 33

[6] A. E. Pérez-Cobas, L. Gomez-Valero, and C. Buchrieser, "Metagenomic approaches in microbial ecology: an update on whole-genome and marker gene sequencing analyses," *Microbial Genomics*, vol. 6, no. 8, aug 2020. 1

[7] N. Sangwan, F. Xia, and J. A. Gilbert, "Recovering complete and draft population genomes from metagenome datasets," *Microbiome*, vol. 4, no. 1, mar 2016. 1

[8] G. V. Uritskiy, J. DiRuggiero, and J. Taylor, "MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis," *Microbiome*, vol. 6, no. 1, sep 2018. 1

[9] D. Gevers, F. M. Cohan, J. G. Lawrence, B. G. Spratt, T. Coenye, E. J. Feil, E. Stackebrandt, Y. V. de Peer, P. Vandamme, F. L. Thompson, and J. Swings, "Re-evaluating

prokaryotic species," *Nature Reviews Microbiology*, vol. 3, no. 9, pp. 733–739, aug 2005. 5

[10] B. A. Adewale, "Will long-read sequencing technologies replace short-read sequencing technologies in the next 10 years?" *African Journal of Laboratory Medicine*, vol. 9, no. 1, nov 2020. 6

[11] B. Berbers, A. Saltykova, C. Garcia-Graells, P. Philipp, F. Arella, K. Marchal, R. Winand, K. Vanneste, N. H. C. Roosens, and S. C. J. D. Keersmaecker, "Combining short and long read sequencing to characterize antimicrobial resistance genes on plasmids applied to an unauthorized genetically modified bacillus," *Scientific Reports*, vol. 10, no. 1, mar 2020. 6

[12] "What is a scaffold?" https://mycocosm.jgi.doe.gov/help/scaffolds.jsf, acessado em 05/2021. 6

[13] S. Nurk, D. Meleshko, A. Korobeynikov, and P. A. Pevzner, "metaSPAdes: a new versatile metagenomic assembler," *Genome Research*, vol. 27, no. 5, pp. 824–834, mar 2017. 7, 31, 50

[14] D. Li, C.-M. Liu, R. Luo, K. Sadakane, and T.-W. Lam, "MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph," *Bioinformatics*, vol. 31, no. 10, pp. 1674–1676, jan 2015. 7, 31, 50

[15] Y. Peng, H. C. M. Leung, S. M. Yiu, and F. Y. L. Chin, "IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth," *Bioinformatics*, vol. 28, no. 11, pp. 1420–1428, apr 2012. 7

[16] D. D. Kang, F. Li, E. Kirton, A. Thomas, R. Egan, H. An, and Z. Wang, "MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies," *PeerJ*, vol. 7, p. e7359, jul 2019. 7

[17] J. Alneberg, B. S. Bjarnason, I. de Bruijn, M. Schirmer, J. Quick, U. Z. Ijaz, L. Lahti, N. J. Loman, A. F. Andersson, and C. Quince, "Binning metagenomic contigs by coverage and composition," *Nature Methods*, vol. 11, no. 11, pp. 1144–1146, sep 2014. 7

[18] Y.-W. Wu, B. A. Simmons, and S. W. Singer, "MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets," *Bioinformatics*, vol. 32, no. 4, pp. 605–607, oct 2015. 7

[19] D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, and G. W. Tyson, "CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes," *Genome Research*, vol. 25, no. 7, pp. 1043–1055, may 2015. 7, 13, 31, 50, 53

[20] J. Hacker and E. Carniel, "Ecological fitness, genomic islands and bacterial pathogenicity," *EMBO reports*, vol. 2, no. 5, pp. 376–381, may 2001. 8

[21] T. Seemann, "Prokka: rapid prokaryotic genome annotation," *Bioinformatics*, vol. 30, no. 14, pp. 2068–2069, mar 2014. 9, 49

[22] T. Tatusova, M. DiCuccio, A. Badretdin, V. Chetvernin, E. P. Nawrocki, L. Zaslavsky, A. Lomsadze, K. D. Pruitt, M. Borodovsky, and J. Ostell, "NCBI prokaryotic genome annotation pipeline," *Nucleic Acids Research*, vol. 44, no. 14, pp. 6614–6624, jun 2016. 9, 29, 40

[23] M. Y. Galperin, Y. I. Wolf, K. S. Makarova, R. V. Alvarez, D. Landsman, and E. V. Koonin, "COG database update: focus on microbial diversity, model organisms, and widespread pathogens," *Nucleic Acids Research*, vol. 49, no. D1, pp. D274–D281, nov 2020. 9, 24

[24] V. Lombard, H. G. Ramulu, E. Drula, P. M. Coutinho, and B. Henrissat, "The carbohydrate-active enzymes database (CAZy) in 2013," *Nucleic Acids Research*, vol. 42, no. D1, pp. D490–D495, nov 2013. 9, 24

[25] L. P. Antunes, L. F. Martins, R. V. Pereira, A. M. Thomas, D. Barbosa, L. N. Lemos, G. M. M. Silva, L. M. S. Moura, G. W. C. Epamino, L. A. Digiampietri, K. C. Lombardi, P. L. Ramos, R. B. Quaggio, J. C. F. de Oliveira, R. C. Pascon, J. B. da Cruz, A. M. da Silva, and J. C. Setubal, "Microbial community structure and dynamics in thermophilic composting viewed through metagenomics and metatranscriptomics," *Scientific Reports*, vol. 6, no. 1, dec 2016. 9

[26] J. Goris, K. T. Konstantinidis, J. A. Klappenbach, T. Coenye, P. Vandamme, and J. M. Tiedje, "DNA–DNA hybridization values and their relationship to whole-genome sequence similarities," *International Journal of Systematic and Evolutionary Microbiology*, vol. 57, no. 1, pp. 81–91, jan 2007. 9

[27] J. C. Setubal, N. F. Almeida, and A. R. Wattam, "Comparative genomics for prokaryotes," in *Comparative Genomics*. Springer New York, dec 2017, pp. 55–78. 10

[28] S. Warrenfeltz, E. Y. Basenko, K. Crouch, O. S. Harb, J. C. Kissinger, D. S. Roos, A. Shanmugasundram, and F. Silva-Franco, "EuPathDB: The eukaryotic pathogen genomics database resource," in *Methods in Molecular Biology*. Springer New York, 2018, pp. 69–113. 11, 26

[29] I.-M. A. Chen, K. Chu, K. Palaniappan, A. Ratner, J. Huang, M. Huntemann, P. Hajek, S. Ritter, N. Varghese, R. Seshadri, S. Roux, T. Woyke, E. A. Eloe-Fadrosh, N. N. Ivanova, and N. C. Kyrpides, "The IMG/m data management and analysis system v.6.0: new tools and advanced capabilities," *Nucleic Acids Research*, vol. 49, no. D1, pp. D751–D763, oct 2020. 11

[30] J. J. Davis, A. R. Wattam, R. K. Aziz, T. Brettin, R. Butler, R. M. Butler, P. Chlenski, N. Conrad, A. Dickerman, E. M. Dietrich, J. L. Gabbard, S. Gerdes, A. Guard, R. W. Kenyon, D. Machi, C. Mao, D. Murphy-Olson, M. Nguyen, E. K. Nordberg, G. J. Olsen, R. D. Olson, J. C. Overbeek, R. Overbeek, B. Parrello, G. D. Pusch, M. Shukla, C. Thomas, M. VanOeffelen, V. Vonstein, A. S. Warren, F. Xia, D. Xie, H. Yoo, and R. Stevens, "The PATRIC bioinformatics resource center: expanding data and analysis capabilities," *Nucleic Acids Research*, oct 2019. 11

[31] L. M. Rodriguez-R, S. Gunturu, W. T. Harvey, R. Rosselló-Mora, J. M. Tiedje, J. R. Cole, and K. T. Konstantinidis, "The microbial genomes atlas (MiGA) webserver: taxonomic and gene diversity analysis of archaea and bacteria at the whole genome level," *Nucleic Acids Research*, vol. 46, no. W1, pp. W282–W288, jun 2018. 12

[32] L.-X. Chen, K. Anantharaman, A. Shaiber, A. M. Eren, and J. F. Banfield, "Accurate and complete genomes from metagenomes," *Genome Research*, vol. 30, no. 3, pp. 315–333, mar 2020. 12

[33] F. Nadalin, F. Vezzi, and A. Policriti, "GapFiller: a de novo assembly approach to fill the gap within paired reads," *BMC Bioinformatics*, vol. 13, no. S14, sep 2012. 12

[34] W. C. Nelson, B. J. Tully, and J. M. Mobberley, "Biases in genome reconstruction from metagenomic data," *PeerJ*, vol. 8, p. e10119, oct 2020. 12, 47

[35] A. Meziti, L. M. Rodriguez-R, J. K. Hatt, A. Peña-Gonzalez, K. Levy, and K. T. Konstantinidis, "The reliability of metagenome-assembled genomes (MAGs) in representing natural populations: Insights from comparing MAGs against isolate genomes derived from the same fecal sample," *Applied and Environmental Microbiology*, vol. 87, no. 6, jan 2021. 13, 45

[36] S. Kurtz, A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S. L. Salzberg, "Versatile and open software for comparing large genomes," *Genome Biology*, vol. 5, no. 2, p. R12, 2004. 17, 29

[37] L. M. Rodriguez-R and K. T. Konstantinidis, "The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes," mar 2016. 20, 29

[38] A. J. Page, C. A. Cummins, M. Hunt, V. K. Wong, S. Reuter, M. T. Holden, M. Fookes, D. Falush, J. A. Keane, and J. Parkhill, "Roary: rapid large-scale prokaryote pan genome analysis," *Bioinformatics*, vol. 31, no. 22, pp. 3691–3693, jul 2015. 20, 29

[39] J. E. Stajich, "The bioperl toolkit: Perl modules for the life sciences," *Genome Research*, vol. 12, no. 10, pp. 1611–1618, oct 2002. 20, 29

[40] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with bowtie 2," *Nature Methods*, vol. 9, no. 4, pp. 357–359, mar 2012. 27, 29, 31

[41] A. Sczyrba, P. Hofmann, P. Belmann, D. Koslicki, S. Janssen, J. Dröge, I. Gregor, S. Majda, J. Fiedler, E. Dahms, A. Bremges, A. Fritz, R. Garrido-Oter, T. S. Jørgensen, N. Shapiro, P. D. Blood, A. Gurevich, Y. Bai, D. Turaev, M. Z. DeMaere, R. Chikhi, N. Nagarajan, C. Quince, F. Meyer, M. Balvočiūtė, L. H. Hansen, S. J. Sørensen, B. K. H. Chia, B. Denis, J. L. Froula, Z. Wang, R. Egan, D. D. Kang, J. J. Cook, C. Deltel, M. Beckstette, C. Lemaitre, P. Peterlongo, G. Rizk, D. Lavenier, Y.-W. Wu, S. W. Singer, C. Jain, M. Strous, H. Klingenberg, P. Meinicke, M. D. Barton, T. Lingner, H.-H. Lin, Y.-C. Liao, G. G. Z. Silva, D. A. Cuevas, R. A. Edwards, S. Saha, V. C. Piro, B. Y. Renard, M. Pop, H.-P. Klenk, M. Göker, N. C. Kyrpides, T. Woyke, J. A. Vorholt, P. Schulze-Lefert, E. M. Rubin, A. E. Darling, T. Rattei, and A. C. McHardy, "Critical assessment of metagenome interpretation—a benchmark of metagenomics software," *Nature Methods*, vol. 14, no. 11, pp. 1063–1071, oct 2017. 27

[42] C. Quince, S. Nurk, S. Raguideau, R. James, O. S. Soyer, J. K. Summers, A. Limasset, A. M. Eren, R. Chikhi, and A. E. Darling, "STRONG: metagenomics strain resolution on assembly graphs," *Genome Biology*, vol. 22, no. 1, jul 2021. 27

[43] D. Merkel, "Docker: lightweight linux containers for consistent development and deployment," *Linux J.*, vol. 2014, no. 239, p. 2, 2014. 29

[44] M. N. Price, P. S. Dehal, and A. P. Arkin, "FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix," *Molecular Biology and Evolution*, vol. 26, no. 7, pp. 1641–1650, apr 2009. 29

[45] H. Zhang, T. Yohe, L. Huang, S. Entwistle, P. Wu, Z. Yang, P. K. Busk, Y. Xu, and Y. Yin, "dbCAN2: a meta server for automated carbohydrate-active enzyme annotation," *Nucleic Acids Research*, vol. 46, no. W1, pp. W95–W101, may 2018. 29

[46] A. Leimbach, "Bac-genomics-scripts: Bovine e. coli mastitis comparative genomics edition," 2016. 29

[47] A. Lafita, S. Bliven, A. Prlić, D. Guzenko, P. W. Rose, A. Bradley, P. Pavan, D. Myers-Turnbull, Y. Valasatava, M. Heuer, M. Larson, S. K. Burley, and J. M. Duarte, "BioJava 5: A community driven open-source bioinformatics library," *PLOS Computational Biology*, vol. 15, no. 2, p. e1006791, feb 2019. 29

[48] P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M. J. L. de Hoon, "Biopython: freely available python tools for computational molecular biology and bioinformatics," *Bioinformatics*, vol. 25, no. 11, pp. 1422–1423, mar 2009. 29

[49] P. Rice, I. Longden, and A. Bleasby, "EMBOSS: The european molecular biology open software suite," *Trends in Genetics*, vol. 16, no. 6, pp. 276–277, jun 2000. 29

[50] P. Danecek, J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies, and H. Li, "Twelve years of SAMtools and BCFtools," *GigaScience*, vol. 10, no. 2, jan 2021. 29

[51] J. G. Lawrence and H. Ochman, "Amelioration of bacterial genomes: Rates of change and exchange," *Journal of Molecular Evolution*, vol. 44, no. 4, pp. 383–397, apr 1997. 46

[52] G. M. Douglas and M. G. I. Langille, "Current and promising approaches to identify horizontal gene transfer events in metagenomes," *Genome Biology and Evolution*, vol. 11, no. 10, pp. 2750–2766, aug 2019. 50