UNIVERSIDADE DE SÃO PAULO

INSTITUTO DE QUÍMICA

PROGRAMA INTERUNIDADES DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

REMIGIO CENEPO ESCOBAR RODRIGUES

RECUPERAÇÃO DE ELEMENTOS GENÉTICOS MÓVEIS E PROSPECÇÃO DE GENES DE RESISTÊNCIA A ANTIBIÓTICOS EM DADOS METAGENÔMICOS DE COMPOSTAGEM TERMOFÍLICA

São Paulo
2023

REMIGIO CENEPO ESCOBAR RODRIGUES

RECUPERAÇÃO DE ELEMENTOS GENÉTICOS MÓVEIS E PROSPECÇÃO DE GENES DE RESISTÊNCIA A ANTIBIÓTICOS EM DADOS METAGENÔMICOS DE COMPOSTAGEM TERMOFÍLICA

Tese apresentada ao Programa Interunidades de Pós-Graduação em Bioinformática da Universidade de São Paulo para a obtenção do título de Doutor em Ciências (Área de concentração: Bioinformática).

Versão corrigida

Orientadora: Profa. Dra. Aline Maria da Silva
Co-orientador: Prof. Dr. João Carlos Setubal

São Paulo
2023

Candidato: Remigio Cenepo Escobar Rodrigues

Título: Recuperação de elementos genéticos móveis e prospecção de genes de resistência a antibióticos em dados metagenômicos de compostagem termofílica.

Tese apresentada ao Programa Interunidades de Pós-Graduação em Bioinformática da Universidade de São Paulo para a obtenção do título de Doutor em Ciências (Área de concentração: Bioinformática).

Aprovado em 27 de outubro de 2023.

Banca Examinadora

Examinador	Prof. Dr. Robson Francisco de Souza Instituto de Ciências Biomédicas Universidade de São Paulo
Examinador	Prof. Dr. Luciano Antonio Digiampietri Escola de Artes, Ciências e Humanidades Universidade de São Paulo
Examinador	Prof. Dr. Julio Cezar Franco de Oliveira DCB-UNIFESP-Diadema
Presidente	Profa. Dra. Aline Maria da Silva Instituto de Química Universidade de São Paulo

DEDICATÓRIA

Dedico este trabalho a minha amada e inesquecível mãe, que se chamava Lari, uma mulher trabalhadora e honesta. Era ela quem me reprimia pela minha desobediência, com o objetivo de me ensinar e mostrar as dificuldades que um dia teria que enfrentar na vida.

Ao meu querido pai Ulices Cenepo Sangama, homem trabalhador, honesto e tranquilo, que desde sua juventude até hoje com 97 anos sempre fica do meu lado em todas as ocasiões.

À minha querida madrinha Ruth da Fonseca, generosa mulher que me deu amor e carinho, foi ela quem me colocou no caminho da educação formal.

À minha querida esposa Marcia, que sempre está me apoiando nas minhas ideias e me ajudando a alcançar os meus sonhos.

Ao meu irmão Marco Antônio que sempre me apoia nos momentos que preciso.

AGRADECIMENTOS

A minha orientadora, Profa. Aline Maria da Silva, por compartilhar comigo, seu tempo, suas ideias, o seu conhecimento e sua experiência ao longo do doutorado.

Ao meu coorientador, Prof. João Carlos Setubal pelas sugestões nas análises bioinformáticas e por garantir acesso aos recursos computacionais para a análise dos dados.

Ao Carlos Morais Piroupo pelo apoio técnico e suporte na instalação de ferramentas e gerenciamento dos recursos computacionais.

Ao Prof. Ronaldo Hashimoto por ter ministrado a aula de reconhecimento de padrões e biologia de sistemas. Foi com ele que tive a oportunidade de conhecer os algoritmos de *machine learning* e grafos de Bruijn.

Ao Prof. Alan Durham por ter ministrado a disciplina de algoritmos em bioinformática. Com ele aprendi os conceitos de HMM e montagem de sequências.

A Profa. Julia Maria Soller pelos ensinamentos na disciplina de Estatística.

Aos Profs. Marie-Anne Van Sluys e Robson F. de Souza pelos ensinamentos sobre elementos genéticos móveis, e pelas sugestões ao trabalho como membros das bancas do EQ, CAC e EPP e Tese.

Aos Profs. Luciano Antonio Digiampietri e Julio Cezar Franco de Oliveira pelas sugestões ao trabalho como membros das bancas do EQ, CAC, EPP e Tese.

Ao meu amigo José Deney Alves de Araújo, aos colegas de laboratório Dra. Layla Martins, Joana Azevedo, Guilhermo Uceda-Campos e Ariosvaldo dos Santos-Junior pelo companheirismo, trocas de ideias e sugestões ao projeto.

Ao Alexandre Sanchez pela recepção e suporte dado no início da pós-graduação.

A todos os meus colegas do IFAM (Instituto Federal do Amazonas) Campus Coari que sempre estão torcendo pelo meu sucesso.

Ao IFAM Campus Coari por permitir o meu afastamento para realização do curso de Doutorado.

Muito obrigado a todos que participaram de forma direta ou indireta da minha formação acadêmica.

Este trabalho foi apoiado por auxílios à pesquisa da CAPES (Projeto 3385/2013) e da FAPESP (Processo 2011/508706) que financiaram a infraestrutura computacional instalada no Laboratório de Bioinformática do IQ-USP.

“*Se enxerguei mais longe,*
foi porque me apoiei sobre os ombros de gigantes”
Sir Isaac Newton

RESUMO

RODRIGUES, R. C. E. Recuperação de elementos genéticos móveis e prospecção de genes de resistência a antibióticos em dados metagenômicos de compostagem termofílica. 2023. 104 páginas. Tese de Doutorado. Programa Interunidades de Pós-Graduação em Bioinformática, Universidade de São Paulo, São Paulo.

A metagenômica *shotgun* permite a caracterização da composição taxonômica e do potencial funcional de microbiomas, independente de isolamento e de cultivo dos microrganismos. A análise funcional de microbiomas é realizada diretamente nas *reads*, nos *contigs* ou em MAGs (*Metagenome Assembled Genome)*. Embora muito utilizada, a análise através de MAGs é limitada visto que os MGEs (*Mobile Genetic Element*) não são geralmente contemplados na montagem. MGEs, como exemplo os plasmídeos, atuam na disseminação de ARGs (*Antibiotic Resistance Gene*) e contribuem para o aumento da resistência antimicrobiana. A recuperação de MGEs e a prospecção de ARGs a partir de dados metagenômicos pode auxiliar na compreensão da função que eles exercem na comunidade microbiana. Neste trabalho, investigamos o mobiloma (conjunto de MGEs), com ênfase no plasmidoma (conjunto de plasmídeos), e o resistoma (conjunto de ARGs) em dados de metagenômica *shotgun* de *short reads* da compostagem termofílica do Parque Zoológico de São Paulo. Esses dados foram obtidos em trabalhos anteriores e possibilitaram elucidar a composição do microbioma da compostagem e realizar sua caracterização funcional, principalmente quanto ao seu potencial de degradação da biomassa lignocelulósica. Entretanto, o conteúdo de MGEs e de ARGs, foi pouco explorado. Para realizar este estudo, estabelecemos uma metodologia de análise para a recuperação dos MGEs e para a prospecção dos ARGs, a partir da integração de distintas ferramentas computacionais e bancos de dados. As *reads* brutas foram submetidas a filtro de qualidade e montadas em *contigs* dos quais ~235 mil *contigs* (8,9% do total) tiveram tamanho ≥ 1 kpb. Desses contigs, recuperamos 24896 *contigs* com 5 ferramentas computacionais de recuperação de plasmídeos. Em seguida, 652 *contigs*, preditos como sendo de plasmídeos por pelo menos duas das ferramentas de recuperação de plasmídeos, foram submetidos a anotação e comparados ao RefSeq de plasmídeos (*NCBI Reference Sequence Database*). Em 649 *contigs* foram encontrados 17112 ORFs que tiveram similaridade contra 3 bases de dados, destas 50% estavam associadas a genes essenciais da base DEG (*Database of Essential Genes*), 77% tiveram similaridade com o RefSeq e 76% com o COG (*The Clusters of Orthologous Genes*). Dentre os *contigs* preditos como plasmídeos, 230 apresentaram características mais robustas de plasmídeos, dos quais apenas uma pequena fração de *contigs* foram similares a 4 plasmídeos encontrados nas bases de dados interrogadas. Nos *contigs* de plasmídeos identificamos uma alta abundância de genes de transposases, de proteínas de replicação e manutenção plasmidial além de mecanismos de defesa incluindo resposta a estresse oxidativo, resistência a metais pesados e genes de resistência a aminoglicosídeos e a sulfonamidas. Cabe destacar que os ARGs identificados estavam mais associados aos *contigs* de cromossomos do que de plasmídeos. Observamos que vários dos *contigs* de plasmídeos estão presentes em mais de uma das amostras coletadas ao longo do processo de compostagem. Verificamos que *reads* do metatranscritoma da compostagem mapearam em genes de transposases e de início de replicação de alguns dos *contigs* de plasmídeos sugerindo que tais elementos estão expressos. Embora ainda existem desafios na montagem de sequências repetitivas que dificultam a recuperação de plasmídeos, a metodologia que estabelecemos possibilitou a recuperação mais precisa de *contigs* de plasmídeos a partir de dados de metagenômica *shotgun* de *short reads*.

Palavras-chave: metagenômica *shotgun*, *short reads*, mobiloma, plasmidoma, resistoma, compostagem.

ABSTRACT

RODRIGUES, R. C. E. Recovery of mobile genetic elements and prospecting for antibiotic resistance genes in metagenomic data from thermophilic composting. 2023. 104 pages. Tese de Doutorado Ph.D. Thesis. Bioinformatics Graduate Program, Universidade de São Paulo, São Paulo.

Shotgun metagenomics allows the characterization of the taxonomic composition and functional potential of microbiomes, independent of the isolation and cultivation of microorganisms. Functional analysis of microbiomes is carried out directly on reads, contigs or MAGs (Metagenome Assembled Genomes). Although widely used, analysis using MAGs is limited because MGEs (Mobile Genetic Elements) are generally not included in the assembly. MGEs, for example plasmids, can disseminate ARGs (Antibiotic Resistance Gene) and contribute to the increase in antimicrobial resistance. The recovery of MGEs and the prospecting of ARGs from metagenomic data can help to understand the role they play in the microbial community. In this work we investigated the mobilome (collection of MGEs), with an emphasis on the plasmidome (collection of plasmids), and the resistome (collection of ARGs) in shotgun metagenomic data from short reads of the thermophilic composting of the São Paulo Zoo Park. These data obtained in a previous work enabled the elucidation of the composition of the composting microbiome and its functional characterization, particularly the potential for degrading lignocellulosic biomass. However, the content of MGEs and ARGs has been little explored. To carry out this study, we established an analysis methodology for recovering MGEs and prospecting ARGs, based on the integration of different computational tools and databases. The raw reads were subjected to a quality filter and assembled into contigs from which ~235 thousand contigs (8.9% of the total) were ≥ 1 kbp in size. Out of these, we recovered 24896 contigs with 5 computational plasmid recovery tools. Next, 652 contigs, predicted to be plasmids by at least two of the plasmid recovery tools, were subjected to annotation and compared to plasmid RefSeq (NCBI Reference Sequence Database). In 649 contigs, 17112 ORFs were found to have similarity against 3 databases, of which 50% were associated with essential genes from the DEG database (Database of Essential Genes), 77% had similarity with RefSeq and 76% with COG (The Clusters of Orthologous Genes). Among the contigs predicted as plasmids, 230 showed more robust plasmid characteristics, of which only a small fraction of contigs were similar to 4 plasmids found in the databases interrogated. In the plasmid contigs we identified a high abundance of genes encoding transposase, plasmid replication and maintenance proteins, as well as defense mechanisms including response to oxidative stress, resistance to heavy metals and resistance to aminoglycosides and sulfonamides. It is worth noting that the ARGs identified were more associated with chromosome contigs than plasmid contigs. We observed that several of the plasmid contigs were present in more than one of the samples collected throughout the composting process. We found that reads from the composting metatranscriptome mapped onto transposase and replication initiation genes of some of the plasmid contigs, suggesting that these elements are expressed. Although there are still challenges in assembling repetitive sequences that make it difficult to recover plasmids, the methodology we established enabled more accurate recovery of plasmid contigs from short reads shotgun metagenomics data.

Keywords: shotgun metagenomics, short reads, mobilome, plasmidome, resistome, composting.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ARO, *Antibiotic Resistance Ontology*

ARDB, *Antibiotic Resistance Genes Database*

AMR, *Antimicrobial Resistance*

ARG, *Antibiotic Resistance Gene*

BWT, *Burrows-Wheeler Transform*

BacMet, *Antibacterial Biocide and Metal Resistance Genes*

BLAST, *Basic Local Alignment Search Too*l

COG, *The Clusters of Orthologous Genes*

CARD, *The Comprehensive Antibiotic Resistance Database*

DEG, *Database of Essential Genes*

ESKAPE, Acrônimo para *Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa* e *Enterobacter spp*.

GC, Guanina e Citosina

GI, *Genomic Island*

HTS, *High-throughput sequencing*

HGT, *Horizontal Gene Transfer*

I-VIP*, Integron Visualization and Identification Pipeline*

IR, *Inverted Repeat*

IS, *Insertion Sequence*

ICE, *Integrative and Conjugative Element*

kpb, kilo pares de bases

k-mer, subsequência de tamanho k de uma sequência de nucleotídeos

L50, Número de *contigs* que cobrem 50% do total dos *contigs* montados

MAG, *Metagenome-Assembled Genome*

MGE, *Mobile Genetic Element*

MBL, *metallo-hydrolase like*

MGR, *Metal Resistance Gene*

MPS*, Marker Protein Sequence*

Mpb, Mega pares de bases

MPF, *Mating Pair Formation*

MDR, *Multidrug-Resistant*

MOB, Conjunto de genes de mobilidade

NCBI, *National Center for Biotechnology Information*

N50, Tamanho do menor *contig* dos *contigs* mais longos que juntos cobrem 50% do total dos *contigs* montados

ORFs, *Open Reading Frame*

oriT, origem de transferência

pb, par de bases

PLACNET, *Plasmid Constellation Network*

PDB, *Protein Data Bank*

RGI, *Resistance Gene Identifier*

RDS, *Replicon Distribution Score*

rRNA, RNA ribossômico

RPKM, *Reads Per Kilobase per Million*

RefSeq, *NCBI Reference Sequence Database*

SMO, *Sequential minimal optimization*

SVM, *Support Vector Machine*

Tn, Transposon

T4CP, *Type IV Coupling Proteins*

T4P, *Type IV Pilus*

T4SS, *Type IV Secretion Systems*

T2SS, *Type II Secretion Systems*

T6SS, *Type VI Secretion Systems*

XDR, *Extensively Drug-Resistant*

ZC4, Composteira 4 do Zoológico

SUMÁRIO

1. INTRODUÇÃO

1.1. Metagenômica *shotgun*

A metagenômica *shotgun* possibilita a caracterização da composição taxonômica e do potencial funcional de comunidades microbianas, independente da necessidade do laborioso cultivo e isolamento de microrganismos. Nessa abordagem o DNA total extraído de amostras ambientais é sequenciado diretamente utilizando-se sequenciamento de alto desempenho (HTS, *High-throughput sequencing*) (Eisen, 2007; Quince *et al*., 2017; Perez-Cobas *et al*., 2020). A composição taxonômica e o potencial funcional são então analisados diretamente no conjunto de sequências não-montadas (*unassembled reads*) ou nos *contigs* obtidos após etapas de montagem das sequências (*assembled reads*). Após a montagem de sequências, os *contigs* que possuem características semelhantes podem ser agrupados (*binning*) e esse conjunto de *contigs* (*bin*) representa, completa ou parcialmente, um genoma recuperado de metagenoma (MAG, *Metagenome-Assembled Genome*), que pode ser classificado taxonomicamente e analisado quanto ao seu potencial funcional para inferências de seu papel ecofisiológico no microbioma em estudo (Chen *et al*., 2020; Perez-Cobas *et al*., 2020).

Atualmente as tecnologias HTS estão baseadas em diferentes abordagens metodológicas que proporcionam a obtenção de sequências curtas de 150 pb a 300 pb (tecnologias de segunda geração como a Illumina e Ion Torrent) ou sequências longas, com mais de 10 kpb, (tecnologias de terceira geração como a Oxford Nanopore e PacBio). Se por um lado as tecnologias que produzem sequências curtas dificultam a análise de elementos repetitivos e variantes estruturais de genomas e metagenomas (Suzuki *et al*., 2019; Xiao & Zhou, 2020), as tecnologias de sequências longas apresentam menor desempenho em termos da quantidade de *reads* geradas e maiores taxas de erro, as quais podem ser eventualmente reduzidas com aplicação de metodologias computacionais de correção de erros (Alic *et al*., 2016; Zhang *et al*., 2020a).

As *short reads* geradas no sequenciamento metagenômico podem ser mapeadas diretamente contra um banco de dados de referência usando ferramentas de bioinformática como Bowtie2 (Langmead & Salzberg, 2012) e BWA (Li & Durbin, 2009) para a pesquisa de sequências similares (Buchfink *et al*., 2015; Ounit *et al*., 2015). No entanto, as *short reads* não fornecem informações suficientes sobre a organização e estrutura genômica. Para contornar essa limitação, são utilizados programas como MetaSpades (Nurk *et al*., 2017) e MEGAHIT (Li *et al*., 2015) que montam as *short reads* em sequências maiores

chamadas de *contigs* que podem ser anotados e analisados quanto sua composição de genes pela comparação contra bancos de dados de genes e proteínas publicamente disponíveis.

Com ferramentas computacionais de *binning* (Yue *et al*., 2020), os *contigs* metagenômicos podem ser agrupados por supostamente pertencerem a um mesmo genoma, constituindo o que é chamado de *bin*. Cada *bin* pode representar um genoma completo ou incompleto que é recuperado a partir do metagenoma e, portanto, denominado de MAG (*Metagenome-Assembled Genome*). Os MAGs são então submetidos a diversas análises, incluindo sua classificação taxonômica. A recuperação de MAGs em metagenomas de diferentes ambientes tem ampliado o repertório de genomas bacterianos conhecidos e permitido inferências sobre a ecofisiologia de organismos não cultivados, os quais representam 95-99% dos organismos preditos como existentes na biosfera (Almeida *et al*., 2021; Nayfach *et al*., 2020; Youngblut *et al*., 2020).

Embora a recuperação de MAGs a partir de *short reads* tenha permitido a adição de uma grande quantidade de genomas aos bancos de dados existentes, ainda não está claro se eles representam a totalidade da diversidade genética nos metagenomas. Avanços recentes em métodos de sequenciamento metagenômico como Hi-C (Burton *et al*., 2014), *read cloud* (Bishara *et al*., 2018), híbrido (Bertrand *et al*., 2019) e *long reads* (Kolmogorov *et al*., 2019) procuraram abordar as deficiências da metagenômica de *short reads* e abriram a possibilidade de que MAGs baseados em *long reads* possam fornecer genomas quase completos (Gounot *et al*., 2022).

A acurácia e completude dos MAGs depende tanto das metodologias de HTS utilizadas na metagenômica *shotgun* como das metodologias computacionais utilizadas na montagem das sequências e *binning* dos *contigs* (Maguire et al., 2020). Embora poderosa e muito utilizada, a análise do potencial funcional de microbiomas através da análise de MAGs tem se revelado um método limitado uma vez que, geralmente, os MAGs geralmente não contemplam os elementos genéticos móveis (MGE, *Mobile Genetic Element*) e ilhas genômicas (GI, *Genomic Island*) (Arredondo-Alonso et al., 2017; Maguire et al., 2020). MGEs e GIs possuem regiões repetitivas e características distintas da sequência do MAG (conteúdo GC, frequência de tetranucleotídeos, entre outras) que interferem na sua inclusão nas etapas de montagem e *binning* de dados metagenômicos de sequências curtas (Maguire *et al*., 2020).

As tecnologias que geram *short reads* e *long reads* possuem suas vantagens e desvantagens exclusivas. Combinar as duas metodologias (abordagem híbrida) pode

aumentar a confiabilidade e a precisão das análises levando a uma maior caracterização dos metagenomas e a obtenção de resultados mais abrangentes. O fato é que as *long reads* podem abranger sequências repetidas e regiões com altos e baixos conteúdo GC e cobrir lacunas nos *contigs* montados com *short reads*, enquanto as *short reads* possuem maior qualidade de sequenciamento e podem ser usadas para corrigir trechos com baixa qualidade de sequenciamento nas *long reads* (Ryan *et al*., 2017).

Gounot *et al*., (2022) analisaram 109 microbiomas do intestino humano e implementaram o método de montagem híbrida proposto por Bertrand *et al*., (2019). Essa estratégia tem exigido aumento em custos de sequenciamento e aumento no custo de computação em nuvem. Por sua vez, tem resultado em um aumento > 61% no número de genomas recuperados versus os genomas recuperados apenas pelo método que emprega a montagem de *short reads*. De modo geral, as montagens híbridas melhoraram consistentemente a recuperação de genomas entre os gêneros, sem viés significativo para nenhum gênero específico, destacando a versatilidade dessa abordagem.

Brown *et al*. (2021) avaliaram o impacto da montagem com *short reads*, *long reads* e montagem híbrida. Eles apontaram a montagem híbrida como uma técnica valiosa para aumentar a confiabilidade e a precisão das análises na contextualização de ARGs e genes vizinhos principalmente aqueles associados a MGEs e marcadores de patógenos em metagenomas ambientais (Brown *et al*., 2021).

1.2. Elementos Genéticos Móveis

MGEs são segmentos de DNA capazes de se moverem dentro de um mesmo genoma, através de mecanismos de transposição, ou entre genomas de células doadoras e receptoras, através de mecanismos de transferência horizontal de genes (HGT, *Horizontal Gene Transfer*), como transformação (absorção de DNA ambiental), conjugação (transferência de DNA por plasmídeo) e transdução (transferência de DNA por bacteriófago) (Frost *et al*., 2005; Juhas *et al*., 2009; Soucy *et al*., 2015; Bertelli *et al*., 2019; Saak *et al*., 2020).

Exemplos de MGEs, são as sequências de inserção (IS, *Insertion Sequence*) ou elementos IS, transposons (Tn), gene cassetes, integrons, plasmídeos, elementos conjugativos e integrativos (ICE, *Integrative and Conjugative Element*) e bacteriófagos. Os MGEs podem ser incorporados em um genoma receptor, formando grupos de genes constituindo ilhas genômicas (Frost *et al*., 2005; Koonin & Wolf, 2008; Juhas *et al*., 2009; Soucy *et al*., 2015; Brito *et al*., 2016; Bertelli *et al*., 2019; Saak *et al*., 2020).

Além de codificarem proteínas que medeiam sua mobilização, os MGEs também podem carregar genes que conferem resistência a antibióticos (ARG, *Antibiotic Resistance Gene*), genes de resistência a metais pesados (MRG, *Metal Resistance Genes*) e fatores de virulência, além de genes que desempenham papéis importantes na adaptação e evolução de bactérias e de comunidades microbianas (Frost *et al*., 2005; Koonin & Wolf, 2008; Brito *et al*., 2016; Saak *et al*., 2020).

O conjunto de todos os MGEs de um genoma ou metagenoma é denominado como mobiloma (Frost *et al*., 2005). O mobiloma contribui para a plasticidade do genoma bacteriano que consiste na capacidade do genoma ser alterável, permitindo a troca fluida de DNA de um microrganismo para outro, isso permite que as bactérias adaptem seus genomas rapidamente para que possam sobreviver a mudanças ambientais e ocupem novos nichos (Tsushima *et al*., 2019), assim fornecendo uma importante fonte de diversidade genética e desempenhando um papel fundamental na ecologia e evolução bacteriana (Rodríguez-Beltrán *et al*., 2020).

As ilhas genômicas (GI) ou ilhas cromossômicas são grandes sequências de DNA especificamente presentes nos genomas de certas cepas bacterianas, mas não nos genomas de suas variantes mais intimamente relacionadas. Eles geralmente estão integrados em um cromossomo bacteriano, mas também podem ser encontrados em plasmídeos ou em fagos que permitem sua mobilização (Darmon & Leach, 2014). Geralmente as ilhas genômicas são adquiridas por HGT. Elas possuem a capacidade de recombinação pela presença de uma integrase, estão associadas com genes tRNAs, e apresentam repetições em ambas as extremidades. Elas codificam uma ampla variedade de características e são nomeadas de acordo com as funções adaptativas que conferem às suas hospedeiras como: ilhas de patogenicidade que codificam um ou mais fatores de virulência, como toxinas, adesinas e invasinas; ilhas metabólicas que conferem a capacidade da bactéria de usar novas fontes de carbono; ilhas de resistência que conferem resistência aos antibióticos; ilhas de simbiose que permite a coevolução da bactéria com sua hospedeira. Ao contrário do genoma central, as ilhas genômicas podem variar muito em sequência e composição de uma espécie bacteriana para outra, e até mesmo dentro de uma espécie. Essa variabilidade e plasticidade permite que as bactérias se adaptem a vários ambientes (Hallstrom *et al*., 2015; Arashida *et al*., 2022).

Os ICEs são encontrados em uma variedade de bactérias, e estão integrados nos cromossomos, mas podem excisar, circularizar e serem transferidos para outras células por conjugação. Além dos principais módulos que medeiam a integração, excisão, conjugação

e regulação, os ICEs codificam uma variedade de funções acessórias, incluindo produção ou inibição de biofilme (Wozniak & Waldor, 2010). Por exemplo o ICE Bs1 encontrado na maioria das cepas de *Bacillus subtilis* confere uma vantagem seletiva, atrasando o desenvolvimento de biofilme e formação de esporos, permitindo que o hospedeiro se dissemine mais do que as células sem ICE Bs1. Isto é vantajoso, porque a hospedeira explora mais o meio ambiente antes da esporulação, aumentando as chances de propagação de ICE Bs1 (Joshua *et al*., 2021).

Os elementos IS são segmentos de DNA relativamente pequenos entre 0,7 a 2,5 kpb. Eles podem ter até duas ORFs (*Open Reading Frame*) que codificam apenas proteínas responsáveis pelas funções envolvidas em sua mobilidade (uma transposase) e são delimitadas por sequências IR (*Inverted Repeat*) terminais curtas (Darmon & Leach, 2014). A inserção de uma IS pode modificar a expressão de alguns genes do hospedeiro. A interrupção de um gene ou de sua sequência reguladora pode levar à inativação do gene. Dependendo do gene inativado, as consequências celulares podem variar de vantajosas a deletérias. Por outro lado, uma IS inserida a montante de um gene pode ativar a expressão deste. Uma IS também pode conter um promotor que pode ativar genes do hospedeiro. Além disso, a inserção de uma IS pode alterar a topologia do DNA no qual está inserido e, às vezes, pode introduzir ou interromper uma sequência reguladora de ligação, afetando a regulação dos genes. É interessante que a transposição de alguns ISs pode ser regulada em certas condições necessárias para ativar a transcrição de operons silenciosos dependendo da presença substrato a ser metabolizado pelas enzimas codificadas no operon (Darmon & Leach, 2014; Hall, 1999).

A transposase é responsável pela catálise da inserção da IS em diferentes sítios de um genoma. As IS são caracterizadas dependendo do tipo de transposase que codificam e devido ao seu mecanismo de cópia (Siguier *et al*., 2015). Em *Acinetobacter baumanii*, a inserção da sequência ISAba-1 sobre o gene *blaAMP-C* com um promotor forte, causa super expressão da cefalosporinase (Heritier *et al*., 2006). Uma IS também está envolvida na mobilização do gene *mrc-1* (*mobile colistin resistance*) no cromossomo de *Actinobacillus pleuropneumoniae* e em diferentes *replicons* de plasmídeos (Snesrud *et al*., 2016).

Transposons são sequências móveis delimitadas por IS. Foram descobertos por volta de 1940 como genes saltadores de milho por Barbara McClintock, a primeira mulher a ganhar sozinha o Prêmio Nobel de Fisiologia e Medicina em 1983, aos 81 anos. Ela também descobriu que, dependendo de onde os Tn são inseridos em um cromossomo, esses elementos móveis podem alterar reversivelmente a expressão de outros genes

(Ravindran, 2012). Existem duas classes de elementos transponíveis que diferem pelo seu mecanismo de mobilização. A primeira classe está baseada em um mecanismo de transcrição reversa de uma molécula de RNA e inserção no local alvo (retrotransposons). A segunda se baseia na excisão do segmento de DNA e inserção do fragmento em outro local do genoma. O mecanismo de inserção no genoma é independente da homologia de sequência, portanto, o transposon pode ser inserido em qualquer sítio do genoma (Beauregard *et al*., 2008; Partridge *et al*., 2018). Transposons foram encontrados associados a ARGs promovendo sua transferência para plasmídeos conjugativos (Frost *et al*., 2005). Por exemplo, os transposons da família Tn3 são disseminadores de ARGs. A Tn3 consiste em uma família grande com representantes em quase todos os filos bacterianos, incluindo proteobactérias, firmicutes e cianobactérias. Os Tn dessa família, possuem três módulos funcionais, o primeiro contém o gene da transposase responsável pela catálise e movimentação do Tn, o segundo contém a resolvase que facilita a replicação do Tn, o terceiro módulo, consiste numa região que pode abrigar diferentes genes acessórios como ARGs e fatores de virulência. Nessa região também podem ser encontrados outros elementos móveis que facilitam ainda mais que o bloco de genes acessórios seja movido para outros locais do cromossomo.

Existem evidências de transposons da família Tn3 que carregam ARGs como é o caso do Tn552 que carrega o gene da β-lactamase (Grindley, 2002; Radstrom *et al*., 1994), e o Tn917 que carrega genes de resistência a macrolídeos. Outros transposons dessa família, foram associados com ARGs que conferem resistência a aminoglicosídeos, vancomicina, tetraciclina, resistência a metais pesados como mercúrio, genes catabólicos de compostos xenobióticos como hidrocarbonetos aromáticos, genes para a degradação de compostos recalcitrantes como tolueno, xilenos e naftaleno que são usados como fonte de carbono pelas bactérias hospedeiras. Outras funções associadas aos transposons da família Tn3 também incluem os fatores de virulência de patógenos. Como é o caso do TnXO1 do plasmídeo pXO1 encontrado no *Bacillus anthracis* que carrega o operon gerX BAC responsável pela germinação de esporos durante a infecção (Nicolas *et al*., 2015).

A principal diferença entre os transposons e elementos IS é que na verdade, os transposons que carregam os genes de mobilização e genes acessórios, são conhecidos como transposons compostos, enquanto as IS são transposons simples pois carregam apenas genes que codificam transposases que catalisam o movimento dos transposons (Tansirichaiya *et al*., 2016).

Genes cassetes são elementos móveis sem replicação autônoma que podem ser encontrados na forma circular livre, são constituídos de ORFs normalmente sem promotor e são flanqueados por regiões chamadas attC que consiste numa sequência conservada de sítio específico de recombinação. Esses locais de recombinação diferem em comprimento e sequência, mas compartilham regiões conservadas em suas extremidades que são geralmente repetições invertidas imperfeitas previstas para formar estruturas de haste-alça. Genes cassetes podem ser encontrados em forma linear como parte de integrons onde podem ser expressos, vários genes cassetes podem ser inseridos no mesmo integron formando uma matriz em tandem, os genes que fazem parte do cassete podem estar associados a várias funções incluindo a resistência a antibióticos e, em muitos casos, a função desses genes ainda é desconhecida (Partridge *et al*., 2009).

Integrons se caracterizam por terem um gene da família intI que codifica normalmente uma tirosina recombinase específica do local (IntI), um local de recombinação attI adjacente e um promotor Pc que controla a expressão do gene cassete (Fluit & Schmitz, 1999). A recombinase tem a capacidade de inserir o gene cassete por recombinação entre os sítios attC x attI, em uma orientação de tal modo que o gene cassete pode ser expresso a partir do promotor Pc. Os integrons foram classificados de acordo com a integrase associada em IntI, Int2, Int3 e Int4 (Escudero *et al*., 2015). A importância dos integrons reside em sua capacidade de recrutar e expressar gene cassetes em sua estrutura. Por sua vez, existem integrons associados a transposons, que permite um certo grau de mobilidade extra entre diferentes locais (Partridge *et al*., 2009).

Os integrons podem ser transferidos entre as bactérias por plasmídeos. Um estudo recente demonstrou, pela primeira vez, que ARGs em integrons da classe 1 podem ser encontrados em fagomídeos que exibem propriedades híbridas de fagos temperados e plasmídeos. A descoberta de integrons da classe 1 em fagomídeos sugere uma rota anteriormente não reconhecida através da qual os fagos podem mediar a disseminação de ARGs associados a genes cassetes (Pfeifer *et al.*, 2022). Outro estudo, também avaliou os integrons da classe 1 em genomas de fagos virulentos que infectam bactérias ambientais, usando ensaios de recombinação attC x attI em *Escherichia coli*, e mostraram que sítios attC transmitidos por fagos podem ser reconhecidos por IntI1 e integrados no sítio attI1. Esses resultados sugerem que os genes cassetes transmitidos por fagos, têm potencial para serem recrutados em integrons bacterianos e vice-versa, o que representa uma rota pouco explorada de HGT entre fagos e bactérias (Qi *et al*., 2023).

Plasmídeos são moléculas de DNA dupla fita extracromossômicos, autoreplicativos, presentes em aproximadamente 50% das bactérias (Clark *et al.*, 2019). Na maioria das vezes são moléculas circulares, embora já tenham sido encontrados plasmídeos lineares em *Borellia*, *Streptomyces*, *Nocardia* e *Rhodococcus* (Dib *et al*., 2015). O conjunto de todos os plasmídeos de uma espécie ou população microbiana é chamado de plasmidoma (Walker, 2012).

Os plasmídeos possuem uma variedade de aplicações na engenharia genética, como vetores para clonagem de sequências de DNA, manipulação e transferência de genes e para expressão de proteínas recombinantes.

De acordo com a capacidade de auto transmissão, os plasmídeos são divididos em três tipos, conjugativo ou auto transmissível, mobilizável e não mobilizável. Essas diferenças ocorrem pela presença ou ausência de genes responsáveis pela conjugação e mobilização. Os plasmídeos conjugativos ou auto transmissíveis são constituídos por quatro módulos: i) uma região de origem de transferência (oriT), ii) genes das proteínas relaxases que constitui o módulo MOB (Conjunto de genes de mobilidade), iii) genes de proteínas de acoplamento do tipo IV (T4CP, *Type IV Coupling Proteins*) e iv) genes do aparato do sistema de secreção do tipo IV (T4SS, *Type IV Secretion Systems*) (Burrus, 2017; Grohmann *et al*., 2018).

Os plasmídeos mobilizáveis contêm apenas o módulo MOB (com ou sem o T4CP) e precisam do (MPF, *Mating Pair Formation*) de um plasmídeo conjugativo para se tornar transmissível por conjugação (Ramsay & Firth, 2017). Os plasmídeos não mobilizáveis por sua vez não possuem nenhum sistema que garanta sua conjugação ou mobilização, mas podem ser transferidos por transformação ou transdução (Smillie *et al*., 2010).

Os plasmídeos possuem características diferentes dos outros MGEs até agora descritos, como a característica de que são elementos autoreplicativos. No entanto, existem os fagomídeos que codificam genes estruturais virais para facilitar a transdução de seus genomas entre hospedeiros, enquanto se replica dentro de bactérias como um plasmídeo (Lobocka, 2004). Os plasmídeos podem ter tamanhos aproximadamente entre 1 kpb a 2 Mpb (Galata *et al*., 2019; Gordon & Christie, 2014). Plasmídeos maiores que 500 kpb são raramente reportados, mas existem relatos de megaplasmídeos em espécies de Alphaproteobacteria (Gordon & Christie, 2014). Isso indica que a quantidade de genes que os plasmídeos podem abrigar em sua estrutura é muito maior do que a de outros MGEs. Existem casos em que diferentes transposons e integrons são encontrados associados com plasmídeos. Essas associações de MGEs facilitam a mobilidade de genes tanto intra,

quanto inter-genômica (Diene & Rolain, 2014). Um caso estudado da evolução dos plasmídeos é a do grupo IncW, que devido à inserção de integrons ou transposons foi capaz de recrutar diferentes genes de resistência a antibióticos (Revilla *et al*., 2008; Partridge *et al*., 2018).

1.3. Genes de resistência a antibióticos

A resistência aos antimicrobianos (AMR, *Antimicrobial Resistance*) é reconhecida pela OMS (Organização Mundial da Saúde) como uma das 10 principais ameaças à saúde pública mundial (WHO, 2020). Um levantamento estimou que vêm ocorrendo de 0,9 a 1,7 milhões de mortes por ano e que esse número pode aumentar para 10 milhões até 2050 (Antimicrobial Resistance Collaborators, 2022).

As infecções adquiridas em hospitais estão entre os principais casos de mortalidade e morbidade, e as crescentes infecções por bactérias Gram-negativas MDR (*Multidrug-Resistant*) e XDR (*Extensively Drug-Resistant*) estão restringindo as opções terapêuticas. Existem relatos que em ambientes hospitalares a taxa de mortalidade por *Pseudomonas aeruginosa* é da ordem de 18 a 61%. Alguns estudos revelaram que *Acinetobacter baumannii* e outros bacilos Gram-negativos não fermentadores estão se tornando MDR. Outros exemplos de bactérias problemáticas incluem o *Staphylococcus aureus* resistente à meticilina (MRSA), o *Mycobacterium tuberculosis* MDR, que é resistente a rifampicina, fluoroquinolona e isoniazida e *Escherichia coli* resistente a colistina e carbapenem, pela aquisição dos ARGs *mcr-1* e *blaNDM-1* (Gandhi *et al*., 2010; Hu *et al*., 2016; Mediavilla *et al*., 2016). O tratamento de infecções associadas a bactérias MDR e XDR ainda é um desafio que leva a problemas significativos para o controle da infecção devido às limitações de agentes antimicrobianos eficazes (Mirzaei *et al*., 2020). O uso indiscriminado de antibióticos exerce uma pressão seletiva para o aumento de bactérias MDR e XDR que estão predominantemente associadas a infecções nosocomiais, particularmente em pacientes imunodeprimidos. Além disso, bactérias MDR têm sido causadoras de infecções adquiridas fora do ambiente hospitalar e, dessa forma, elevando o risco de transmissão na população (van Duin & Paterson, 2016).

Os ARGs conferem às bactérias mecanismos de resistência aos antibióticos, os quais são agrupados segundo seu modo de ação tais como:

i) modificação ou degradação do antibiótico, tais como a hidrólise do anel β-lactâmico catalisada por β-lactamases que ocasionam perda da função de antibióticos β-lactâmicos como penicilina e carbapenem (Queenan & Bush,

2007) ou a acetilação por transferases que inativam aminoglicosídeos (Garneau-Tsodikova & Labby, 2016);

ii) alteração do sítio alvo do antibiótico impedindo a interação com a molécula alvo. Um exemplo é a metilação do rRNA que impede a união de aminoglicosídeos à subunidade 16S do ribossomo (Gutierrez *et al*., 2012);

iii) proteção do sítio alvo do antibiótico. A resistência a fluoroquinolonas mediadas pelos genes *qnr* é um exemplo. As proteínas codificadas por esses genes são pentapeptídeos repetidos que se associam a DNA girase e a topoisomerase IV impedindo a ligação do antibiótico e gerando níveis moderados de resistência (Vetting *et al*., 2011);

iv) expressão de variantes de enzimas com baixa afinidade pelo antibiótico, mas que cumprem as mesmas funções das enzimas originais. Como exemplo pode ser mencionado a resistência a sulfonamidas através das variantes de di-hidrofolato redutases *sul1* e *sul2* (Skold, 2001);

v) regulação da expressão de genes, permitindo às bactérias suportarem baixas concentrações de antibióticos por tempos curtos. Acredita-se que esta resposta é sincronizada com outros mecanismos como a indução da resposta SOS. Isso induz mudança na estabilidade genética da bactéria que permite explorar fenótipos de maior resistência (Handel *et al*., 2014);

vi) limitação da capacidade de entrada do antibiótico devido a alteração em propriedades da superfície bacteriana;

vii) aumento da capacidade de efluxo do antibiótico pela expressão de bombas de efluxo de diferentes classes de antibióticos, contribuindo para o fenótipo MDR (Reygaert, 2018).

A presença e abundância de ARGs em diferentes ambientes têm sido exploradas. Por exemplo, a abundância de ARGs varia entre 10^6 - 10^{11} cópias/g de peso seco (sólido) no esterco de gado, ou 10^6 - 10^{12} cópias/mL em águas residuais (He *et al*., 2020), valores comparativamente maiores do que os detectados no solo ou fontes de água (Ahmed *et al*., 2023). Além dos ARGs, os MGEs também são amplamente encontrados no esterco animal, o que aumenta o risco de disseminação dos ARGs quando estão associados a MGEs. Estudos relataram que a abundância de MGEs no esterco animal variou de 10^6 a 10^{15} cópias/g de peso seco (Cheng *et al*., 2021; Fan *et al*., 2020). Solos enriquecidos com esterco e irrigados com águas residuais são importantes reservatórios de ARGs (Forsberg

et al., 2012; Cytryn, 2013; Forsberg *et al*., 2014; McKinney *et al*., 2018; Van Goethem *et al*., 2018).

Outro importante reservatório de ARGs é o próprio microbioma intestinal humano (Penders *et al*., 2013). Há evidência de que a alta densidade microbiana nesses ecossistemas facilite a transferência horizontal de ARGs, contribuindo para a disseminação da AMR (Penders *et al*., 2013; von Wintersdorff *et al*., 2016; Sitaraman, 2018). Em um estudo metagenômico da microbiota intestinal de humanos e animais de consumo como bovinos, aves e suínos, foi detectado pela primeira vez na América Latina, ARGs que codificam carbapenemases, como *bla*AIM-1, *bla*CAM-1, *bla*GIM-2 e *bla*HMB-1. Esse estudo é um exemplo da importância da metagenômica como uma ferramenta para rastrear a colonização de bactérias resistentes em animais produtores de alimentos e humanos (Carvalho *et al*., 2022). O conjunto de ARGs em um genoma ou metagenoma é denominado de resistoma.

Infelizmente a disseminação da resistência bacteriana aos antibióticos é desproporcional à descoberta de novos antimicrobianos. A indústria farmacêutica vem fazendo pouco investimento ou até mesmo se retirando da pesquisa e do desenvolvimento de novos antibióticos por vários motivos, incluindo os altos custos dos ensaios clínicos e a falta de incentivos financeiros. Além disso, o modelo de negócios mudou, pois qualquer novo antibiótico ficará restrito ao uso emergencial nos hospitais, que reduz drasticamente o potencial de lucro (Hegemann, 2023).

1.4. Recuperação de MGEs em dados metagenômicos

A recuperação de MGEs é usualmente realizada usando métodos laboratoriais laboriosos e dependentes de cultivo microbiano, os quais não fornecem uma visão abrangente do mobiloma (Dib *et al*., 2015). Como alternativa a esses métodos laboratoriais, foram desenvolvidas estratégias computacionais capazes de recuperar sequências de MGEs a partir de dados metagenômicos (Jorgensen *et al*., 2014). Entretanto, a recuperação de MGEs a partir de dados metagenômicos *shotgun* é desafiador, principalmente quando se trata de plasmídeos. Primeiro, porque a proporção de *reads* correspondentes ao DNA do cromossomo bacteriano é maior em relação ao DNA plasmidial (Dib *et al*., 2015). Segundo, porque as repetições que ocorrem nos plasmídeos complicam o processo de montagem pela formação de quimeras e *contigs* curtos e prejudicam o processo de recuperação, principalmente quando as amostras são sequenciadas com tecnologias HTS que geram sequências curtas (Antipov *et al*., 2019). Terceiro, os cromídeos e os

megaplasmídeos (tamanho > 330 kpb) possuem genes essenciais e genes do sistema de replicação plasmidial, o que complica a distinção entre plasmideo e cromossomo (Harrison *et al*., 2010). Quarto, os plasmídeos podem se integrar ao cromossomo bacteriano, tornando difícil caracterizar computacionalmente os *contigs* como plasmídeos ou cromossomos (Harrison *et al*., 2010).

As dificuldades associadas à identificação e caracterização de plasmídeos podem explicar parcialmente sua escassez nos bancos de dados como ocorre por exemplo no RefSeq versão 203 que em 2020 contava com 63.237 genomas bacterianos, e apenas 5.222 sequências de plasmídeos (Stockdale, 2022). Com isso, percebemos que ainda não existe um conceito computacionalmente bem definido para distinguir um plasmídeo de um cromossomo bacteriano (Rozov *et al*., 2017).

Existem poucos programas especializados para a recuperação de plasmídeos de dados metagenômicos e, por essa razão, são então utilizados programas destinados à recuperação de plasmídeos de genomas isolados. Os algoritmos de recuperação de plasmídeos usam principalmente duas abordagens, a primeira consiste na classificação binária dos *contigs*, baseado na comparação contra uma sequência de referência, os *contigs* podem ser fragmentados em k-mer para depois serem comparados, essa classificação prediz a origem do *contig*, isto é, se é de cromossomo bacteriano ou de plasmídeo. A segunda abordagem consiste na reconstrução do plasmídeo que também pode se basear numa referência ou em grafos montados (montagem de novo) (Paganini *et al*., 2021). A Figura 1 mostra as principais ferramentas de recuperação de plasmídeos.

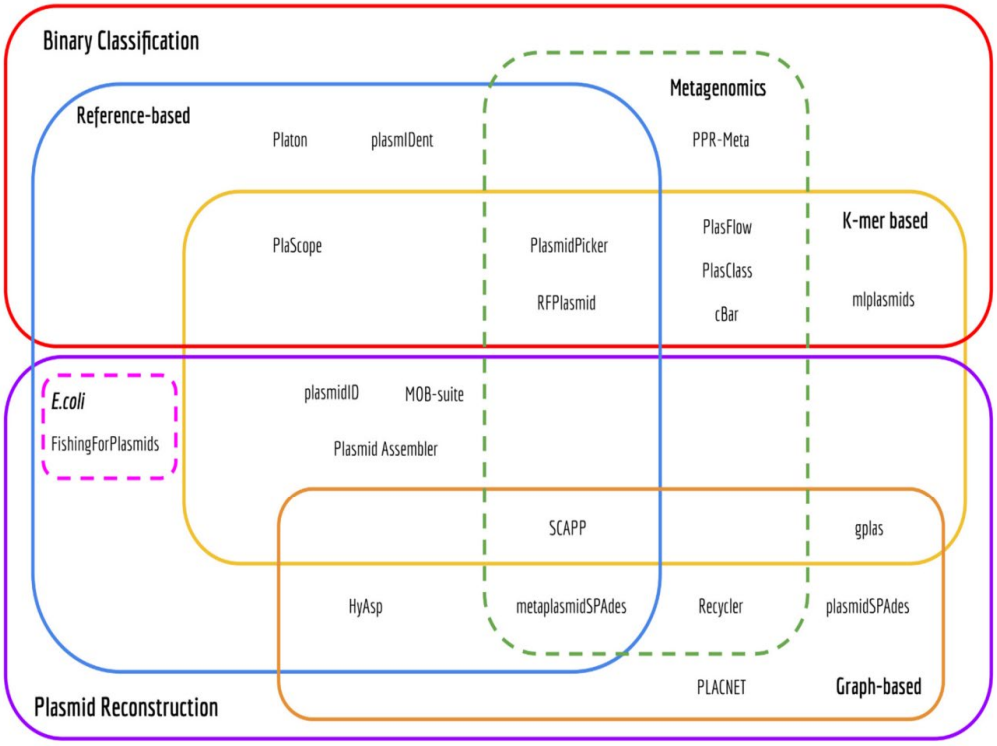


Figura 1: Diagrama de Euler mostrando as principais ferramentas de bioinformática para a predição de plasmídeos disponíveis até 2020. Fonte: Paganini *et al.*, 2021.

Entre os programas especializados estão o MetaPlasmidSpades (Antipov *et al*., 2019), Recycler (Rozov *et al*., 2017), Plasflow (Krawczyk *et al*., 2018), cBar (Zhou & Xu, 2010), SCAPP (Pellow *et al*., 2021), PlasmidPicker (a ferramenta está disponível em https://github.com/haradama/PlasmidPicker), RFPlasmid (Linda *et al*., 2020), PlasClass (Pellow, Mizrahi & Shamir, 2020) e PPR-Meta (Fang *et al*., 2019). O MetaPlasmidSpades e o Recycler são capazes de recuperar *contigs* circulares a partir de grafos de Bruijn montados com MetaSpades (Nurk *et al*., 2017). Embora essas ferramentas tenham revelado uma série de novos plasmídeos, são relatados muitos falsos positivos, especialmente em situações em que a cobertura cromossômica não é uniforme (Antipov *et al*., 2019). O Plasflow e o cBar usam assinaturas genômicas e técnicas de aprendizagem de máquina (Arredondo-Alonso *et al*., 2018). O Plasflow usa um modelo treinado em rede neural para distinguir a porcentagem do conteúdo GC entre o plasmídeo e o cromossomo do hospedeiro (Krawczyk *et al*., 2018). O cBar usa um algoritmo de otimização mínima sequencial (SMO, *Sequential Minimal Optimization*) para distinguir a frequência de pentâmeros (Zhou & Xu, 2010).

O RFPlasmid usa um classificador *Random Forest* treinado com uma abordagem híbrida, identificando genes marcadores cromossômicos e plasmidiais usando banco de dados de proteínas e frequências de pentâmeros. No entanto, é limitado para apenas 17

espécies bacterianas de diferentes táxons nos quais os classificadores foram treinados e conta com um modelo genérico para os organismos desconhecidos para os dados de metagenômica (Linda *et al*., 2020).

O PPR-Meta permite a identificação simultânea de fragmentos de fagos e plasmídeos de metagenomas usando uma Rede Neural Convolucional. Em vez de frequências k-mer, esta ferramenta usa matrizes *one-hot* para representar sequências de nucleotídeos e aminoácidos.

O PlaScope e o PlasmidPicker executam pesquisas k-mer em bancos de dados de plasmídeos de referência. Essa estratégia é muito rápida, mas limitada a detectar k-mers que estão presentes no banco de dados subjacente.

Semelhante ao cBAR, O mlplasmids também depende das frequências do pentâmero, mas usa um modelo de *Support Vector Machine* (SVM) para determinar a origem dos *contigs* para uma única espécie e contém modelos para *Escherichia coli*, *Klebsiella pneumoniae* e *Enterococcus faecium* (Arredondo-Alons *et al*., 2018).

HyAsP e SCAPP usam uma abordagem híbrida, baseados em referência e de novo. HyAsP foi desenvolvido para predizer plasmídeos em genomas isolados e se baseia na identificação de uma alta densidade de genes plasmidiais conhecidos em banco de dados e leva em consideração a uniformidade do conteúdo GC bem como a alta cobertura das reads (Müller & Chauve, 2019). O SCAPP, por outro lado, é projetado para encontrar plasmídeos em montagens de metagenoma. Esse algoritmo inicia encontrando possíveis *contigs* de plasmídeo com base em duas estratégias: (1) a busca de genes específicos de plasmídeo usando um banco de dados curado e (2) atribuindo peso a cada *contig* com base na saída de PlasClass, um classificador binário baseado em aprendizagem de máquina. O gráfico de montagem é então consultado para encontrar caminhos cíclicos de cobertura uniforme, semelhante ao Recycler, mas priorizando a inclusão de *contigs* com forte evidência de origem plasmidial (Pellow *et al*., 2021).

O Platon é um *pipeline* que além de recuperar plasmídeos também prediz genes de resistência a antibióticos e usa uma variedade de programas como Prodigal (Hyatt *et al*., 2010), Nucmer do pacote MUMmer (Kurtz *et al*., 2004), BLAST+, DIAMOND (Buchfink *et al*., 2015) e banco de dados incluindo o RefSeq de plasmídeos (Brooks *et al*., 2019) e de cromossomos (Tatusova *et al*., 2015), PlasmidFinder e UniProt UniRef90 (Apweiler *et al*., 2004), além de uma base de dados chamada *Marker Protein Sequences* (MPS) criada pelos próprios autores do *pipeline*. A principal diferença em relação ao OriTFinder é que o Platon usa um novo sistema de métrica que foi chamado pelos autores de *Replicon Distribution*

Score (RDS) e que usa uma abordagem estatística para distinguir se um gene pertence a um plasmídeo ou a um cromossomo.

Outros programas que não estão indicados na Figura 1 também são usados para a recuperação de plasmídeos a partir de genomas isolados, entre esses o PlasmidFinder (Carattoli *et al*., 2014) e o OriTfinder (Li *et al*., 2018). O PlasmidFinder consiste em um banco de dados de grupos de incompatibilidade de plasmídeos e usa o BLASTn para buscar homologia entre esses grupos (Carattoli *et al*., 2014). Sua principal limitação é a abrangência do banco dados que é composto principalmente por *replicons* de plasmídeos que possuem como hospedeiras as bactérias da família Enterobacteriaceae, o que é um fator limitante para sua empregabilidade em estudos metagenômicos (Krawczyk *et al*., 2018). A abordagem de pesquisa de similaridade em banco de dados é ampliada no PLACNET (*Plasmid Constellation Network*) que também usa o BLAST (*Basic Local Alignment Search Tool*) para comparar as sequências contra um banco de dados de referência e, em seguida, usa uma análise de rede para reconstruir os plasmídeos (Lanza *et al*., 2014). No entanto, este programa conta com a curadoria manual dos *clusters* de sequências obtidos, evitando, assim, seu uso em qualquer pipeline de anotação automática. Além disso, os resultados obtidos do PLACNET não são totalmente reproduzíveis e dependem da experiência do usuário (Schwengers *et al*., 2020).

O OritFinder se baseia na busca por similaridade de regiões como OriT, genes de virulência, resistência a antibióticos, relaxases, sistemas T4CP, T4SS e grupos de incompatibilidade (Li *et al*., 2018; Schwengers *et al*., 2020). O OriTfinder faz essa busca contra um banco de dados chamado OriTDB (Li *et al*., 2018) usando HMMer (Finn *et al*., 2011) e variações de BLAST. Sua principal limitação é que os *contigs* precisam ser analisados de um em um, algo que também ocorre no PlasmidFinder.

Cada uma dessas abordagens possui vantagens e desvantagens. Por exemplo, as abordagens baseadas em variações de cobertura do sequenciamento são incapazes de detectar plasmídeos com número de cópias iguais ao do cromossomo, enquanto os programas baseados em k-mers, tendem a identificar apenas plasmídeos já conhecidos (Schwengers *et al*., 2020). Isso frequentemente leva a distintos perfis em termos de sensibilidade e especificidade, que muitas vezes são tendenciosos para uma das métricas. Como isso impacta na análise conduzida, uma escolha deve ser feita entre classificações conservadoras ou mais agressivas (Arredondo-Alonso *et al*., 2017).

1.5. Prospecção de ARGs em dados metagenômicos

A prospecção de ARGs em dados genômicos e metagenômicos tem se beneficiado da existência de banco de dados de ARGs como o CARD (*The Comprehensive Antibiotic Resistance Database*) (Alcock *et al*., 2020) e o ARDB (*Antibiotic Resistance Genes Database*) (Liu & Pop, 2009), os quais foram desenvolvidos a partir das análises em fontes como o Genbank (Benson *et al*., 2018) e o PDB (*Protein Data Bank*) (Zou *et al*., 2015). A pesquisa por homologia nestes bancos é feita com uso de ferramentas de bioinformática como BLAST (Altschul *et al*., 1990), DIAMOND (Buchfink *et al*., 2015), Bowtie2 (Langmead & Salzberg, 2012) e RGI (*Resistance Gene Identifier*) (McArthur *et al*., 2013).

O CARD é um banco de dados curado atualizado mensalmente e associado ao ARO (*Antibiotic Resistance Ontology*) que descreve os mecanismos de resistência dos ARGs, seus alvos moleculares, suas mutações e os fenótipos associados. Para a análise de ARGs, conta com os algoritmos BLAST e RGI. A primeira versão do CARD foi desenvolvida em 2013 (McArthur *et al*., 2013) com o propósito de reunir e organizar todos os ARGs conhecidos que atuam em diferentes classes de antibióticos. Esses ARGs eram principalmente encontrados em genomas de patógenos isolados a partir de casos clínicos. No entanto, nos últimos anos ARGs encontrados em metagenomas de diversos ambientes também passaram a ser incluídos no CARD (Alcock *et al*., 2020).

Em 2018 foi publicado o DeepARG (Arango-Argoty *et al*., 2018), uma ferramenta baseada em técnicas de *deep leanning* para a prospecção de novos ARGs em dados metagenômicos e que usa *features* de ARGs obtidas a partir de um banco de dados curado chamado DeepARG-DB que reúne informações de ARGs dos bancos de dados CARD, ARDB e UNIPROT. Em 2019 os autores do DeepARG desenvolveram um *pipeline* chamado nanoARG com interface web (Arango-Argoty *et al*., 2019) que integra o deepARG como algoritmo de prospecção de ARGs além de outros módulos: i) para a prospecção de genes de resistência a metais pesados (MGRs, *Metal Resistance Genes*) que utiliza o DIAMOND para a análise de similaridade de proteínas contra um banco de dados curado chamado BacMet (*Antibacterial Biocide and Metal Resistance Genes*); ii) para a prospecção de genes associados a MGEs como transposase, integrase e recombinase usando análise de similaridade também com o DIAMOND e I-VIP (*Integron Visualization and Identification Pipeline*) contra o banco de dados não-redundante do NCBI (*National Center for Biotechnology Information*); iii) para a classificação taxonômica dos *contigs* usando a ferramenta Centrifuge, um classificador metagenômico rápido e preciso que usa BWT (*Burrows-Wheeler Transform*); iv) para a classificação de patógenos incluindo os

pertencentes ao grupo ESKAPE (*Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa* e espécies de Enterobacter) que são responsáveis por uma porcentagem substancial de infecções nosocomiais difíceis de tratar. Além disso, o nanoARG permite ver a co-ocorrência entre ARGs, MRGs, MGEs e patógenos bacterianos (Arango-Argoty *et al*., 2019).

1.6. Compostagem termofílica

A compostagem é uma técnica usada para acelerar a decomposição da matéria orgânica mediada por microrganismos (Jurado *et al*., 2014). No processo da compostagem a matéria orgânica sólida biodegradável é convertida em "composto" estável do tipo húmus que pode ser manuseado, armazenado e utilizado como biofertilizante. A decomposição bioxidativa que ocorre nesse processo é realizada por uma complexa comunidade de microrganismos (microbiota), cuja composição e estrutura variam dependendo de fatores como temperatura, pH, aeração, umidade e características dos substratos orgânicos (Insam & de Bertoldi 2007). O metabolismo microbiano impulsiona alterações de pH e o rápido aumento da temperatura, que pode atingir até 80ºC, dependendo do tipo de compostagem (Gajalakshmi & Abbasi, 2008; Lopez-Gonzalez *et al*., 2013; Lopez-Gonzalez *et al*., 2015). O estudo do microbioma da compostagem tem sido importante tanto para elucidar as vias de degradação de biomassa lignocelulósica como para a descoberta de novos microrganismos, novos genomas, novos genes e novas vias metabólicas, demonstrando a surpreendente diversidade microbiológica e genética associada a compostagem (Partanen *et al*., 2010; Martins *et al*., 2013; Jurado *et al*., 2014; Lopez-Gonzalez *et al*., 2015; Antunes *et al*., 2016; Amgarten *et al*., 2017; Braga *et al*., 2021; Guima, 2021).

A compostagem é considerada uma alternativa eficaz e bastante econômica para o tratamento de resíduos orgânicos, inclusive com a eliminação de patógenos, os quais não sobrevivem a fase termofílica do processo (Gajalakshmi & Abbasi, 2008). Durante esse processo, observa-se a redução de ARGs e MGEs pré-existentes na matéria orgânica utilizada como substrato (Wang *et al*., 2017; Zhang *et al*., 2017; Ezzariai *et al*., 2018; Gou *et al*., 2018; Liao *et al*., 2018; Zhou *et al*., 2019; Zhang *et al*., 2020b). Porém, dependendo das condições de compostagem, alguns tipos de ARGs podem persistir (Zhang *et al*., 2015), o que representa risco de disseminação para o meio ambiente. Por exemplo, ARGs (*acrB*, *qnr*S2, *bla*CTX-M-1 e *van*C) e MGEs (intl2, tnpA-6) podem ser removidos, enquanto outros, por exemplo *erm*F, *sul*1, *tet*X, *aad*A, intl3, IS26 e IS1247, persistiram ou se recuperaram durante a fase de resfriamento (Cao *et al*., 2020). Uma das possíveis razões por trás do

enriquecimento desses genes pode ser a remoção parcial de potenciais bactérias hospedeiras, porque mudanças na dinâmica bacteriana foram consideradas como o principal fator biótico para a redução de ARGs (Du *et al*., 2019). Bactérias dos filos Proteobacteria e Actinobacteria são consideradas uma importante fonte de ARGs, exibem resistência a múltiplos antibióticos e sua abundância foi observada na fase de maturidade do processo de compostagem (Ahmed *et al*., 2023).

O resistoma da compostagem tem sido investigado através de qPCR (Zhou *et al*., 2019) e abordagens metagenômicas (Su *et al*., 2015; Xie *et al*., 2016; Wang *et al*., 2017; Liao *et al*., 2018; Zhang *et al*., 2020b). Entretanto, ainda são poucos os estudos com análises abrangentes e integradas do repertório completo de ARGs e MGEs e da co-ocorrência de ARGs com MGEs na compostagem.

O Parque Zoológico de São Paulo (São Paulo, Brasil) utiliza a compostagem para tratamento dos resíduos orgânicos como esterco dos animais, restos de alimentação dos animais, e resíduos de podas de árvores e dos jardins do parque, que inclui um fragmento de Mata Atlântica Trata-se de uma compostagem termofílica que tem duração de aproximadamente 90 dias para produção do composto maduro, o qual é usado como fertilizante nas hortas que produzem os vegetais para alimentação dos animais (Martins *et al*., 2013; Ramos *et al*., 2019). Análises utilizando dados de metagenômica *shotgun* revelaram a enorme diversidade microbiana da compostagem do Parque Zoológico de São Paulo e seu potencial funcional, principalmente aquele relacionado à degradação da biomassa lignocelulósica, tanto a partir de sequências não-montadas como a partir de *contigs* e MAGs (Martins *et al*., 2013; Antunes *et al*., 2016; Lemos *et al*., 2017; Braga *et al*., 2021; Guima, 2021). Porém, o conteúdo de MGEs e seu potencial funcional, incluindo a presença de ARGs, foram pouco explorados nestes conjuntos de dados metagenômicos.

2. OBJETIVOS

Estabelecer uma metodologia de análise para identificar o repertório de MGEs, principalmente plasmídeos, e de ARGs na compostagem termofílica do Parque Zoológico de São Paulo utilizando-se distintas ferramentas computacionais para a recuperação e identificação de MGEs e de ARGs a partir de dados de metagenômica *shotgun* de *short reads*.

3. MATERIAIS E MÉTODOS

3.1. Conjunto de dados metagenômicos de amostras da compostagem

Os conjuntos de dados de metagenômica e de metatranscritômica utilizados neste trabalho são de amostras coletadas ao longo do processo de compostagem de uma composteira do Parque Zoológico de São Paulo denominada ZC4 e foram obtidos em trabalho anterior de nosso grupo (Antunes *et al.*, 2016).

A composteira ZC4 foi montada em 05/08/2013 e foram coletadas amostras nos dias 1, 3, 7, 15, 30, 64, 67, 78 e 99. No dia 63 a pilha foi revirada para aeração. O processo de compostagem de ZC4 apresentou um perfil termofílico, com temperaturas entre 60ºC e 75ºC até o dia 78, e ~50ºC no dia 99 (Antunes *et al.*, 2016). O DNA total dessas amostras foi extraído e submetido ao sequenciamento *short reads* na plataforma Illumina MiSeq (formato de 500 ciclos com *reads* de extremidades pareadas (PE) com tamanho de 250 pb). O RNA total dessas mesmas amostras também foi extraído (exceto da amostra do dia 67) e sequenciado nas plataformas Illumina MiSeq e Illumina HiSeq2500 (formato de 200 ciclos com *reads* de extremidades pareadas (PE) com tamanho máximo de 100 pb). Os procedimentos de coleta, processamento destas amostras e de sequenciamento na plataforma Illumina estão detalhados em Antunes *et al.*, 2016. Os metagenomas e metatranscritomas das amostras de ZC4 foram explorados em estudos anteriores quanto à composição, estrutura e potencial funcional da comunidade bacteriana a partir de *contigs* e de MAGs bacterianos recuperados (Antunes *et al*., 2016; Braga *et al*., 2021).

3.2. Ferramentas de bioinformática e bancos de dados

As ferramentas de bioinformática e bancos de dados utilizados para análises de MGEs e ARGs nos metagenomas da compostagem ZC4 estão listados na Tabela 1. A descrição suscinta dessas ferramentas está apresentada adiante.

Tabela 1: Ferramentas computacionais utilizadas nas análises de MGEs e ARGs

Ferramenta/versão	Função
Fastqc 0.11.9; Multiqc 1.12	Análise de qualidade das *reads*
Trimommatic 0.39	Trimagem de *reads*
MetaSpades 3.15.3	Montagem de *contigs*
Metaquast 5.1.0	Análise de qualidade de *contigs*
Reformat/BBMap 38.94	Filtragem de *contigs* por tamanho
MetaPlasmidSpades 3.14.1	Recuperação de *contigs* de plasmídeos

Plasflow 1.1	Recuperação de *contigs* de plasmídeos
Recycler 0.7	Recuperação de *contigs* de plasmídeos
Platon 1.6	Recuperação de *contigs* de plasmídeos
Scapp 0.1.4	Recuperação de *contigs* de plasmídeos
Barrnap 0.9	Busca genes de rRNA
Filterbyname/BBMap 38.94	Remoção de *contigs* com genes de rRNA
Cd-hit-est 4.8.1	Agrupamento de *contigs*
MetaGeneMark 3.26	Busca de ORFs
Blastp	Anotação
Nucmer/MuMmer 4.0	Análise de similaridade
Bowtie2 2.5.1	Mapeamento de *reads*
Isescan 1.7.2.3	Busca de IS e Tn
Integron_finder 1.5.1	Busca de integron
RGI-bwt 6.0.0	Busca de ARGs por mapeamento de *reads*
RGI 6.0.2	Busca de ARGs em *contigs*
RefSeq 2022	Banco de dados de sequências de aminoácidos e de nucleotídeos de plasmídeos
COG 2020	Banco de dados de genes ortólogos de procariotos
DEG 2020	Banco de dados de genes essenciais de procariotos
CARD 2021	Banco de dados de ARGs

3.2.1. Análise de qualidade das reads

A qualidade das sequências (*reads)* foi analisada com os programas FASTQC (Andrews, 2010) e Multiqc (Ewels *et al*., 2016). A trimagem das sequências foi realizada com Trimommatic (Bolger *et al*., 2014), com o parâmetro: PE R1.fastq R2.fastq -baseout R.fastq HEADCROP:15 SLIDINGWINDOW:5:20 MINLEN:50 CROP:235. Somente as *reads* com qualidade ≥ 20 e tamanho ≥ 50 pb foram mantidas para as próximas análises.

3.2.2. Montagem de novo

O conjunto de *reads* de cada amostra foi montado separadamente em *contigs* com MetaSpades (Nurk *et al*., 2017), com o parâmetro: --meta -k 21, 33, 55, 77, 99, 113, 117, 121, 127 -1 R1.fastq -2 R2.fastq (-k corresponde ao tamanho do kmer). A qualidade da montagem foi avaliada com Metaquast (Mikheenko *et al*., 2016). Após a montagem e análise de qualidade da montagem, foi realizada a separação dos *contigs* por tamanho usando o programa reformat do pacote BBMap (Bushnell, 2014). Foram obtidos dois conjuntos de *contigs*, o primeiro contendo *contigs* com tamanho de 300 pb a < 1 kpb, e o segundo contendo *contigs* ≥ 1 kpb.

3.2.3. Recuperação de *contigs* de plasmídeos

A recuperação de *contigs* potencialmente de plasmídeos foi realizada somente nos *contigs* ≥ 1 kpb utilizando-se cinco ferramentas diferentes:

i) MetaPlasmidSpades (Antipov *et al*., 2019), com o parâmetro: --plasmid -k 21, 33, 55, 77, 99, 113, 117, 121, 127 -1 R1.fastq -2 R2.fastq. É importante destacar que o programa MetaPlasmidSpades faz uma nova montagem a partir das *reads* filtradas. *Contigs* < 1 kpb foram descartados;

ii) Plasflow (Krawczyk *et al*., 2018), com o parâmetro: --threshold 0.9 (equivale a 0.9 de probabilidade de uma sequência ser de plasmídeo);

iii) Recycler (Rozov *et al*., 2017), com o parâmetro: -g assembly-graph.fastg -k 55 -b alinhamento.sort.bam. Foi usado como entrada um arquivo .fastg contendo grafos montados com MetaSpades e um arquivo .bam contendo o alinhamento ordenado das *reads* contra os *contigs* montados. O arquivo .bam foi preparado com o auxílio do programa BWA (Li e Durbin, 2009) e Samtools (Li *et al*., 2009);

iv) Platon versão 1.6 (Schwengers *et al*., 2020), com parâmetros padrão;

v) Scapp (Pellow *et al*. 2021), com parâmetros padrão.

3.2.4. Busca de genes de rRNA

Nos *contigs* recuperados pelas 5 ferramentas de recuperação de plasmídeos foi realizada uma busca por genes de rRNA com a ferramenta Barrnap (Seemann, 2013), com parâmetros padrão. Os *contigs* que tinham pelo menos um gene de rRNA foram removidos do conjunto de *contigs* recuperados e foram considerados como *contigs* de cromossomos. Removemos esses *contigs* porque levamos em consideração que genes de rRNA não são normalmente encontrados em plasmídeos. Entretanto, já foi relatado em *Aureimonas* sp. um operon de rRNA em um pequeno plasmídeo com um alto número de cópias (Anda *et al*., 2015). A remoção foi realizada com o programa filterbyname do pacote BBMap (Bushnell, 2020).

3.2.5. Agrupamento de *contigs* recuperados por mais de uma ferramenta

Os potenciais *contigs* de plasmídeos recuperados por mais de uma ferramenta, foram agrupados com o programa cd-hit-est (Weizhong & Adam, 2006). Como argumento, foi fornecido o valor 1 para o parâmetro -c, o que permitiu agrupar somente os *contigs* com

100% de identidade, independentemente do tamanho. Apenas um *contig* de cada grupo e recuperado pelo menos com duas ferramentas foi selecionado. Nos grupos onde havia *contigs* com tamanhos diferentes, foi selecionado o *contig* maior. Os grupos contendo apenas um *contig*, foram desconsiderados, uma vez que os *contigs* desses grupos não foram recuperados por mais de uma ferramenta sendo portanto removidos do conjunto de *contigs* recuperados e foram considerados como pertencentes a sequências de cromossomos.

3.2.6. Anotação

Os *contigs* recuperados com mais de uma ferramenta e sem genes de rRNA, foram submetidos ao programa MetaGeneMark (Zhu *et al*., 2010) para a identificação de ORFs. Para verificar a existência de genes essenciais, as ORFs foram comparadas contra as proteínas de bactérias do banco de dados DEG (*Database of Essential Genes*) (Luo *et al*., 2014) atualizado em 2020. A anotação da ORFs foi realizada por meio da comparação contra as proteínas não-redundantes de plasmídeos do RefSeq (NCBI *Reference Sequence Database*) seguido de uma comparação adicional para verificar a categoria funcional das ORFs contra o banco de dados COG (*The Clusters of Orthologous Genes*) (Galperin *et al*., 2021) atualizado em 2020. As comparações foram realizadas com Blastp (Altschul *et al*., 2009). Em todas as comparações foi usado como critério de similaridade e-value < 1e-6 e bitscore > 50 conforme sugerido por Pearson (2013). A análise foi realizada com auxílio de *script* em Python versão 3.9 (Van & Drake, 2009).

3.2.7. Busca de plasmídeos conhecidos com Nucmer

Os *contigs* que permaneceram no conjunto de *contigs* recuperados após o processo de anotação foram submetidos a um alinhamento global, usando o programa Nucmer do pacote MUMmer (Kurtz *et al*., 2004) com parâmetros padrão, contra o banco de dados RefSeq de nucleotídeos de plasmídeos (Brooks *et al*., 2019). Foram considerados como *contigs* de plasmídeos conhecidos aqueles que tinham identidade e cobertura ≥ 90%. Os *contigs* sem similaridade aceitável, foram considerados como potenciais plasmídeos desconhecidos.

3.2.8. Busca de plasmídeos conhecidos com Bowtie2

Para verificar a existência de plasmídeos conhecidos nos metagenomas, as *reads* foram mapeadas contra o banco de dados RefSeq de nucleotídeos de plasmídeos usando Bowtie2 (Langmead & Salzberg, 2012), com o parâmetro: -a --no-unal --sensitive. As sequências com cobertura de mapeamento ≥ 90% foram consideradas como pertencentes a plasmídeos conhecidos. O resultado foi comparado com o resultado da análise de similaridade descrita acima com Nucmer para verificar plasmídeos já conhecidos entre os *contigs*.

3.2.9. Busca de ISs, Transposons e Integrons

As buscas por outros elementos genéticos móveis como ISs, Transposons e Integrons foram realizadas nos *contigs* de plasmídeos conhecidos e desconhecidos, assim como também no conjunto de *contigs* considerados como sendo de cromossomos. As ferramentas usadas na busca foram: Isescan (Xie & Tang, 2017) para IS, e Transposons e Integron_finder (Cury *et al*., 2016) para Integrons, ambas com parâmetros padrão.

3.2.10. Prospecção de ARGs

A prospecção de ARGs foi realizada nas *reads* e em todos os *contigs* ≥ 300 pb, o que inclui os *contigs* de plasmídeos como também os *contigs* de cromossomos. Para isso, foram usados os algoritmos RGI-bwt (Guitor *et al*., 2019) e RGI respectivamente integrado ao banco de dados CARD (Alcock *et al*., 2020).

3.2.11. Verificação de *contigs* de plasmídeos repetidos em mais de uma amostra

Para verificar se um *contig* de plasmídeo se repete em mais de uma amostra (dia) da composteira ZC4, foi realizado o mapeamento das *reads* de todas as amostras contra os *contigs* de plasmídeos recuperados. Para isso foi usado o programa Bowtie2 (Langmead & Salzberg, 2012). O *contig* com 100% de cobertura em mais de uma amostra foi considerado como *contig* que se repete nos dias da compostagem.

3.2.12. Análise de genes expressos em *contigs* de plasmídeos

As *reads* do metatranscritoma da compostagem (amostras da série temporal da composteira ZC4) foram submetidas aos programas FASTQC (Andrews, 2010) e

Trimommatic (Bolger *et al*., 2014) para análise de qualidade e trimagem. *Reads* com qualidade ≥ 20 e comprimento ≥ 50 pb foram mapeadas usando Bowtie2 contra as ORFs identificadas com o programa MetaGeneMark nos *contigs* de plasmídeos conhecidos e desconhecidos. As ORFs que tiveram cobertura de mapeamento ≥ 90%, foram consideradas como genes expressos, sendo calculado o valor de RPKM *(Reads per kilo base per million mapped reads*) de cada gene para cada amostra usando a fórmula (Scholz, 2022):

RPKM = numReads/(geneLength/1000*totalNumReads/1000000)

numReads - número de *reads* mapeados no gene
geneLength - tamanho do gene
totalNumReads - total de *reads* mapeadas da amostra

As ORFs consideradas expressas, foram anotadas com blastp contra o RefSeq de proteínas de plasmídeos. Essa etapa permitiu apontar os genes de *contigs* de plasmídeos preditos como expressos ao longo do processo de compostagem.

4. RESULTADOS

4.1. Fluxograma de análise dos dados de metagenômica e metatranscritômica

Para identificar e caracterizar o repertório de MGEs, principalmente plasmídeos, e de ARGs na compostagem termofílica do Parque Zoológico de São Paulo, estabelecemos o fluxograma apresentado na Figura 2 para análise de *reads* de metagenômica *shotgun* e de metatranscritômica. Este fluxograma foi definido após diversos testes que não estão descritos aqui e que possibilitaram ajustes e melhorias na recuperação e identificação dos MGEs, com ênfase nos plasmídeos.

4.2. Processamento das *reads* de metagenômica *shotgun* da compostagem

4.2.1. Análise de qualidade de *reads* e trimagem

Na etapa de análise de qualidade e trimagem das *reads* brutas de metagenômica *shotgun*, observamos mais de 10% de diferença na proporção de bases AT e CG nas 14 bases iniciais e nas 6 bases finais. Esta é uma anomalia característica da metodologia de

sequenciamento, uma vez que o esperado é que ao longo da *read* a proporção de AT e GC seja concordante. Observamos em algumas *reads* das amostras até 0,6% da base N (base indeterminada) em diferentes posições ao longo das *reads*, com uma maior frequência no início da *read*. Não foi detectada nenhuma *read* super-representada > 1%. Foi detectado uma baixa presença do adaptador transposase Nextera (< 1%) em algumas *reads* das amostras, principalmente na extremidade final das *reads*. O tamanho das *reads* variou entre 37-251 pb, sendo que a maioria das *reads* tinha o tamanho entre 247-251 pb. A média de qualidade das *reads* variou entre 2-38 (Phred *score*), sendo que *reads* com qualidade > 20 eram as mais abundantes.

Após a trimagem, a diferença maior do que 10% na proporção de bases AT e GC não foi mais observada. A porcentagem da base N passou a ser 0, o mesmo ocorreu com os adaptadores. O tamanho das *reads* mudou para 52-235 pb, com a maioria ficando com o tamanho entre 232-235 pb. A média de qualidade das *reads* passou a ser maior que 23 Phred Score, com uma grande frequência de *reads* com qualidade entre 28-38 Phred Score. A Figura 3 mostra a proporção de AT e GC ao longo das *reads*, a distribuição de tamanho das *reads* e a média de qualidade das *reads* antes e após o processo de trimagem.

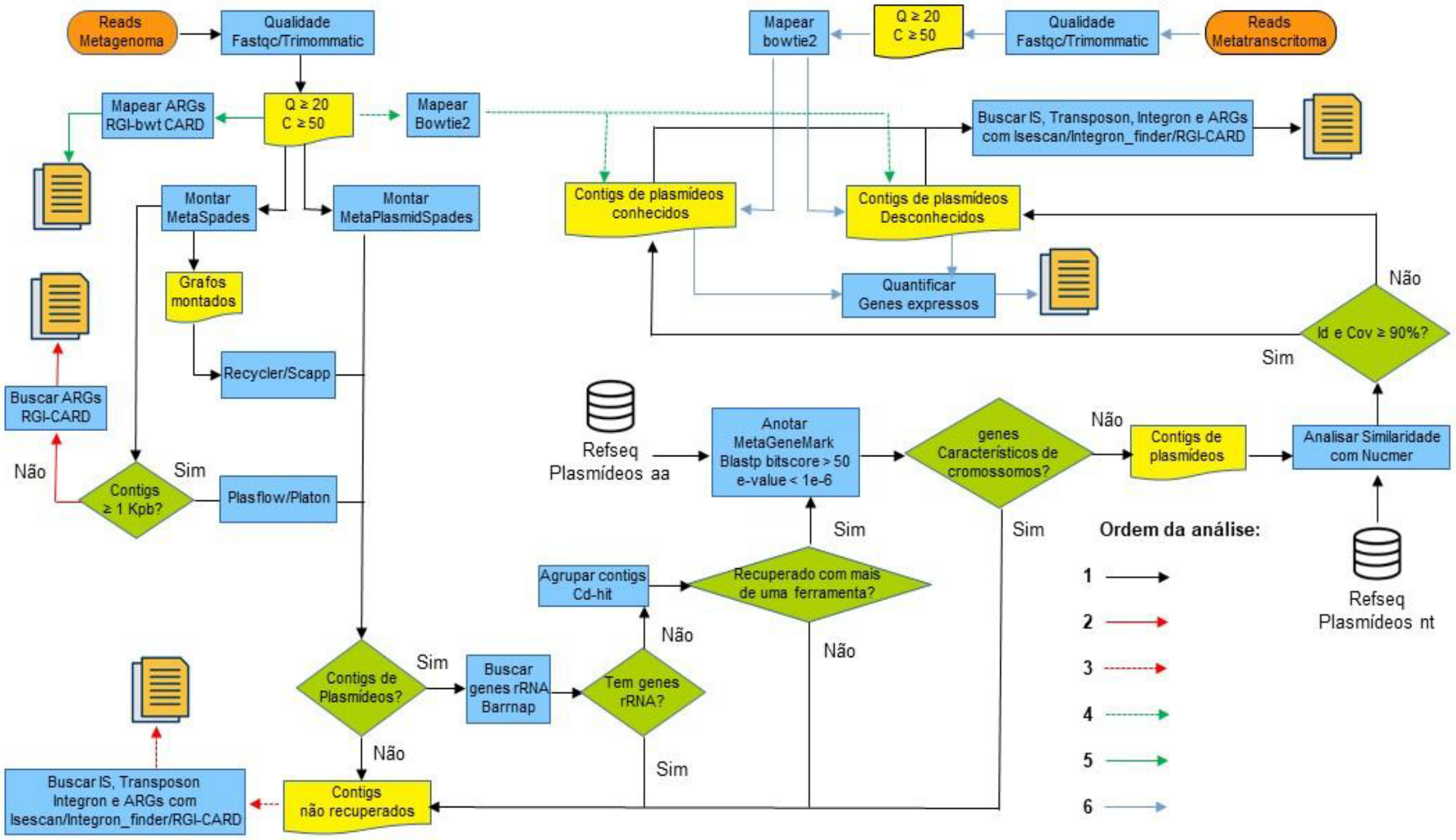


Figura 2: Fluxograma de análise dos dados de metagenômica e metatranscritômica.

As análises foram realizadas na ordem indicada pelas setas em diferentes cores. Abreviações indicadas na figura: Q = Qualidade; C = Comprimento; Id = Identidade: Cov = Cobertura: aa = aminoácidos: nt = nucleotídeos.

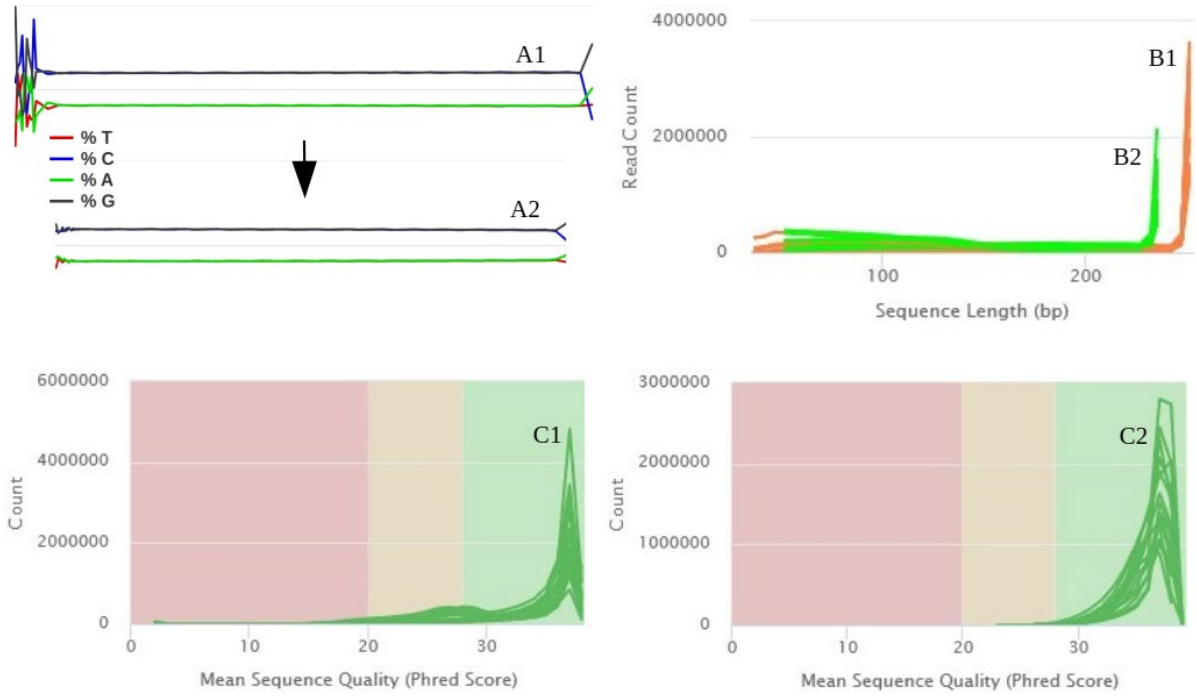


Figura 3: Diferenças observadas nas amostras entre as *reads* brutas e as *reads* trimadas.
A1 – existência de anomalia na proporção de AT e GC nas extremidades das *reads*. A2 – ausência da anomalia após trimagem; B1 – (em laranja) tamanho inicial das *reads*; B2 – (em verde) tamanho das *reads* após trimagem; C1 – média da qualidade inicial das *reads*; C2 – média da qualidade das *reads* após trimagem. Os gráficos foram gerados com o MultiQC (Ewels *et al*., 2016).

O conjunto de *reads* brutas pareadas totalizou 56,3 milhões de *reads*, variando de 4,1 milhões (amostra do dia 01) até 11,2 milhões (amostra do dia 78) (Antunes *et al*., 2016). A porcentagem das *reads* duplicadas foi de até 3,5%. A média do conteúdo GC variou entre 58-64% e a média do tamanho das *reads* variou entre 150-229 pb.

Após o processo de trimagem, restaram 46,6 milhões das *reads* pareadas, o que corresponde a 82,7% do conjunto inicial. O valor mínimo e máximo por amostra ficou em 3,5 milhões (amostra do dia 67) e 8,2 milhões de *reads* (amostra do dia 78), respectivamente. Houve diminuição das *reads* duplicadas em todos os dias. Não houve variação na média do conteúdo GC em relação ao observado para as *reads* brutas. O tamanho médio das *reads* diminuiu e variou entre 125-186 pb. A estatística geral das *reads* brutas e trimadas estão apresentadas na Tabela 2.

A Figura 4 mostra a variação do conteúdo GC nas *reads* das amostras da série temporal. No dia 01 e no dia 64, observa-se uma variação maior do conteúdo GC tanto nas *reads* brutas como nas *reads* trimadas. Possivelmente isso reflete a maior variedade e abundância de microrganismos mesofílicos, coincidente com a menor temperatura verificada (abaixo de 55ºC) nestas amostras (Castaldi *et al*., 2005; Antunes *et al* 2016; Ramos *et al*., 2019).

Tabela 2: Estatística geral das *reads* brutas e *reads* trimadas.

	***Reads* brutas**				***Reads* trimadas**			
Dia/*Reads*	% Dups	% GC	Tamanho	Seqs	% Dups	% GC	Tamanho	Seqs
01/R1	1,0	59	206	4,1	0,8	59	186	3,7
01/R2	0,9	59	206	4,1	0,8	59	180	3,7
03/R1	1,8	62	229	4,7	1,5	62	163	4,0
03/R2	1,5	63	229	4,7	1,4	62	152	4,0
07/R1	3,4	61	179	4,6	2,9	60	162	3,7
07/R2	2,8	61	181	4,6	2,6	60	151	3,7
15/R1	2,8	62	175	7,2	2,4	62	134	5,9
15/R2	2,6	62	178	7,2	2,3	62	125	5,9
30/R1	2,0	63	197	4,8	1,7	63	153	4,1
30/R2	1,7	63	199	4,8	1,6	62	144	4,1
64/R1	1,5	58	205	7,2	1,3	58	182	6,5
64/R2	1,4	59	206	7,2	1,2	57	176	6,5
67/R1	1,2	61	186	4,2	1,0	60	167	3,5
67/R2	1,0	61	188	4,2	0,9	60	156	3,5
78/R1	3,5	63	150	11,2	2,7	63	138	8,2
78/R2	3,3	63	154	11,2	2,5	63	127	8,2
99/R1	1,6	64	171	8,3	1,3	64	146	6,9
99/R2	1,5	64	171	8,3	1,2	64	141	6,9

R1 corresponde as *reads forward;* R2 corresponde as *reads reverse*; Dups – é a porcentagem das *reads* duplicadas; GC é a porcentagem do conteúdo GC; Tamanho é a média do tamanho das *reads* (nt); Seqs – é a quantidade de *reads* em milhões.

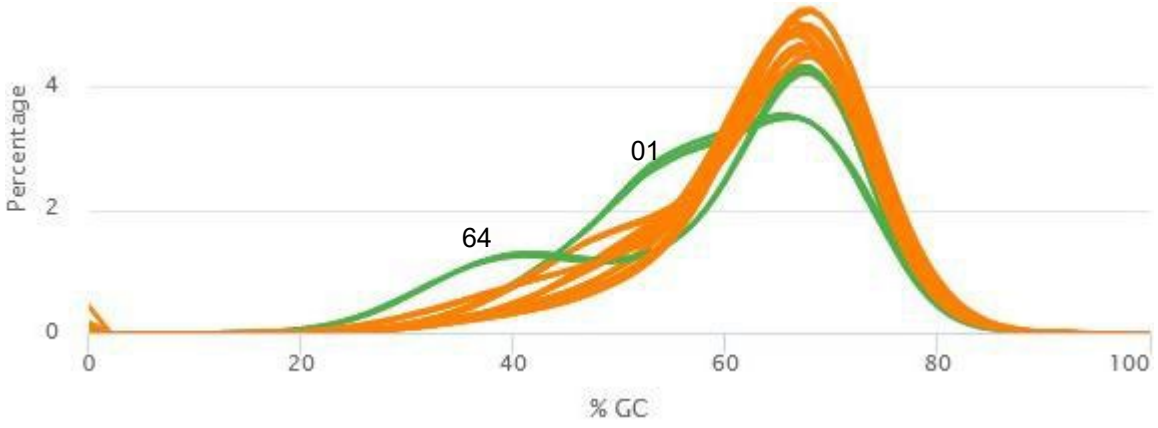


Figura 4: Conteúdo GC nas *reads* das amostras da série temporal da compostagem.
As amostras dos dias 01 e 64 que apresentaram maior variação do conteúdo GC estão destacadas em verde. O Gráfico foi gerado com o MultiQC (Ewels *et al*., 2016).

4.2.2. Montagem de *contigs*

Foram montados aproximadamente 2,6 milhões de *contigs* a partir das sequências metagenômicas de todas as amostras da série temporal da composteira ZC4 utilizando-se o programa MetaSpades (Nurk *et al.*, 2017). A menor quantidade de *contigs* montados ocorreu no dia 07 com 181,71 mil, e a maior quantidade ocorreu no dia 64, com 512,89 mil (Tabela 2, Figura 5A). A profundidade média de cobertura na maior parte dos *contigs* foi < 4x (Figura 5B). Foi observado que em todas as amostras, os *contigs* com tamanho menor que 1 kpb (128 pb a 149 kpb) foram mais abundantes (91,1%) e apenas ~235 mil tiveram o tamanho ≥ 1 kpb, o que equivale a 8,9% do total dos *contigs* montados (Figura 5C). A média do conteúdo GC variou entre 58-65% nas amostras. Com um pico de 68%. Da mesma forma como ocorreu com o conteúdo GC das *reads*, nos *contigs*, a variação do conteúdo GC também foi maior nos dias 01 e 64 mostrados com as linhas vermelho claro e vermelho escuro respectivamente na Figura 5D.

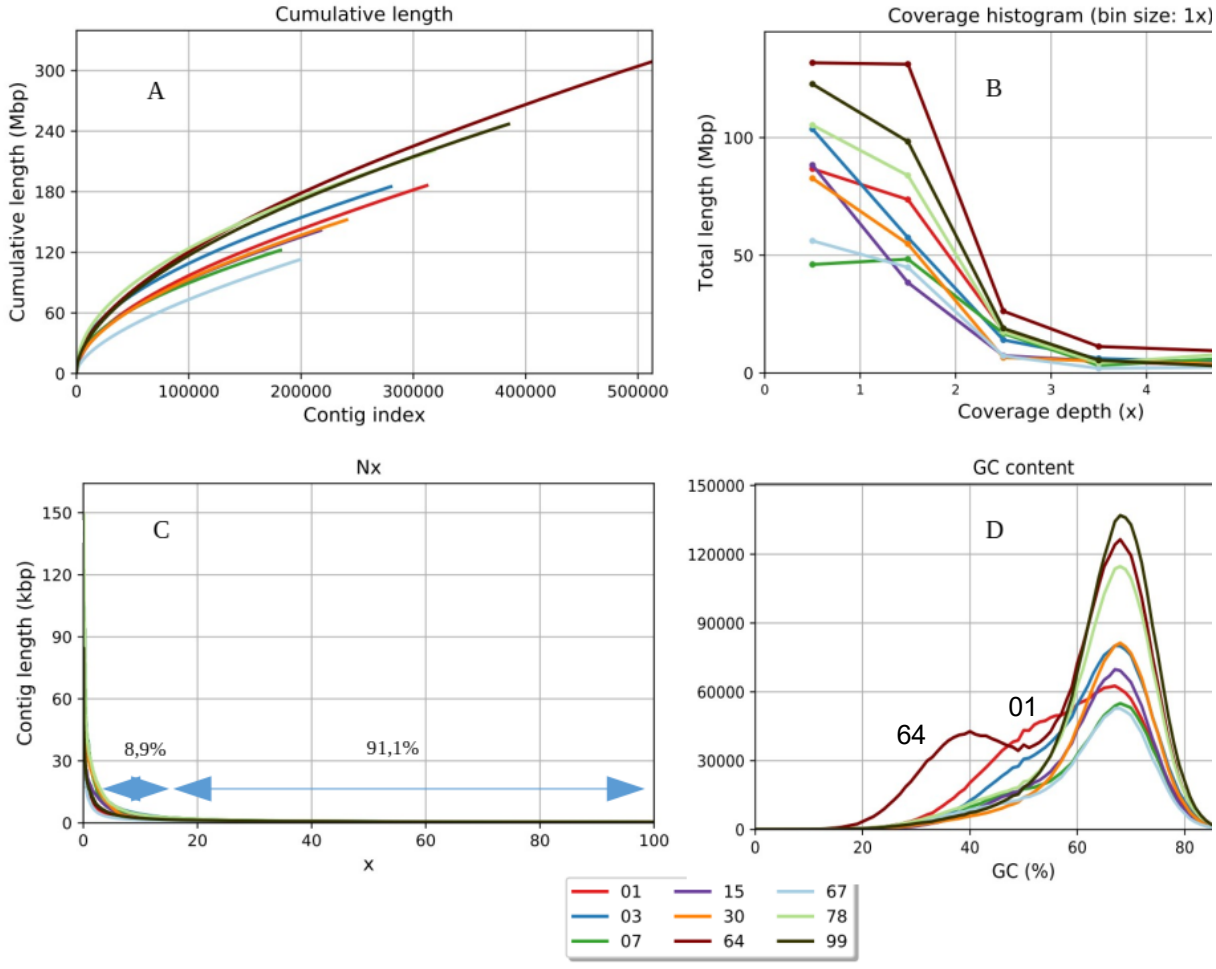


Figura 5: Montagem dos *contigs* a partir das *reads* de cada amostra da série temporal da compostagem. Quantidade (A), profundidade média de cobertura (B) tamanho (C) e média do conteúdo GC (D) (dias 01, 03, 07, 15, 30, 64, 67, 78 e 99). O tamanho cumulativo dos *contigs* está indicado no eixo y dos gráficos em A e B. Os gráficos foram gerados com o Quast (Gurevich *et al.*, 2013).

O valor de N50 (tamanho do menor *contig* dos *contigs* mais longos que juntos cobrem 50% do total dos *contigs* que foram montados em cada amostra) e L50 (número de *contigs* que cobrem 50% do total dos *contigs* montados em cada amostra) foi avaliado para verificarmos a porcentagem de fragmentação (L50%) do conjunto de *contigs* em cada amostra. Quanto maior o valor de L50% (L50% = (L50*100)/Total de *contigs)* e menor o tamanho de N50, mais fragmentados é o conjunto de *contigs*. A tabela 3 mostra que o menor tamanho de N50 (526 pb) e o L50% é 32,6% foi verificado na montagem da amostra do dia 67. No dia 78 podemos observar que o N50 é de 678 pb e seu L50 é 24,8%. Em resumo, os resultados indicam que *contigs* de maior tamanho foram montados a partir das *reads* do dia 78 e os de menor tamanho são do dia 67.

Tabela 3: N50 e L50 dos *contigs* montados por amostra da compostagem.

Dia	Número de *contigs*	maior *contig* (pb)	N50 (pb)	L50	L50 (%)
67	199336	41050	526	65020	32,6
01	312072	54692	553	92800	29,7
64	512890	50357	567	154869	30,2
30	240662	93873	603	68864	28,6
99	384799	84524	623	110268	28,7
03	280081	75769	625	71513	25,5
15	217441	48830	629	60139	27,7
07	181713	106795	643	45727	25,2
78	318599	149225	678	78977	24,8

As amostras 67 e 78 estão destacadas para indicar a ocorrência da maior e da menor fragmentação, respectivamente.

4.3. Recuperação de potenciais *contigs* de plasmídeos

A recuperação de potenciais *contigs* de plasmídeos foi realizada com 5 ferramentas distintas, que usam diferentes abordagens. A Tabela 4 mostra a quantidade de *contigs* recuperados pelas distintas ferramentas.

Tabela 4: Quantidade de *contigs* recuperados por ferramenta.

Dia	Plasflow	MetaPlasmidSpades	Platon	Recycler	Scapp
01	1908	488	97	6	5
03	2074	736	37	11	7
07	1115	467	10	13	5
15	1581	470	20	15	8
30	1611	509	5	10	3
64	3610	1079	66	30	19
67	800	302	19	16	7
78	2804	867	10	8	4
99	3244	777	11	8	4
Total					**24896**

Ao todo foram recuperados 24896 *contigs* como sendo de plasmídeos. A maior quantidade foi recuperada com Plasflow seguido de MetaPlasmidSpades. Scapp foi a ferramenta que recuperou a menor quantidade de *contigs*. Em relação ao tamanho dos *contigs* recuperados, MetaPlasmidSpades foi a ferramenta que recuperou os maiores *contigs* a partir dos dados de todos os dias da compostagem, sendo o maior de 395 kpb recuperado no dia 67 (Figura 6). É importante destacar que o MetaPlasmidSpades faz uma nova montagem, o que explica a recuperação de *contigs* maiores do que os obtidos com outras ferramentas que utilizaram a montagem realizada com MetaSpades descrita no item 4.2.2. O maior *contig* recuperado pelo Plasflow foi de 93,8 kpb no dia 30. Com o Platon, o maior foi de 84 kpb no dia 99. Para Recycler e Scapp, o tamanho dos *contigs* não passou de 50 kpb (Figura 6).

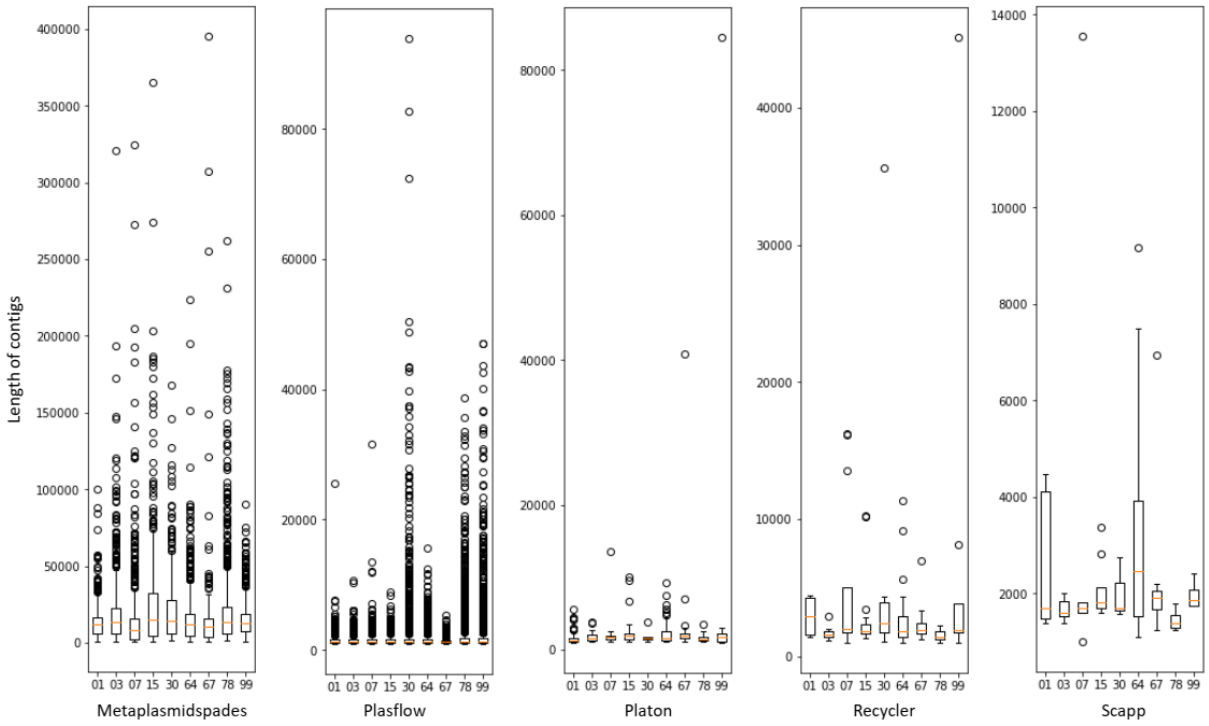


Figura 6: Tamanho (pb) dos *contigs* recuperados pelas 5 ferramentas que foram usadas.

4.3.1. Identificação de genes de rRNA em *contigs* recuperados

Ao todo foram identificados 417 genes de rRNA em 301 *contigs*. Foram encontrados os genes dos RNA 5S, 16S e 23S nas quantidades 140, 141 e 136 respectivamente. Em 88 *contigs* havia mais de 1 gene de rRNA e em 23 *contigs* estavam presentes os três genes simultaneamente. A maior quantidade de *contigs* contendo genes de rRNA foram aqueles recuperados com MetaPlasmidSpades (366 genes), seguido pelo Plasflow (48 genes). Apenas 3 *contigs* recuperados com Recycler tinham esses genes e nas demais ferramentas

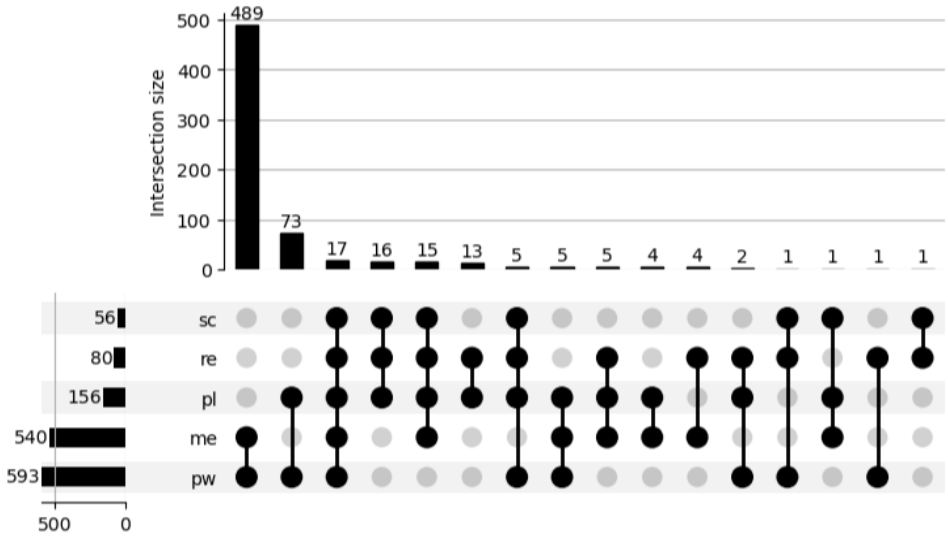
não houve nenhum caso positivo (Tabela 5). Os *contigs* contendo genes de rRNA foram removidos do conjunto de *contigs* recuperados restando ainda 24595 *contigs*.

Tabela 5: Quantidade de genes de rRNA encontrados em *contigs* recuperados como sendo de plasmídeos.

Ferramenta	Quantidade de genes rRNA
MetaPlasmidSpades	366
Plasflow	48
Recycler	3
Platon	0
Scapp	0
Total	**417**

4.3.2. Agrupamento de *contigs* recuperados

Utilizamos o cd-hit-est para agrupar os *contigs* 100% idênticos, resultando em 23280 grupos. Apenas 652 grupos continham *contigs* recuperados por mais de uma ferramenta. De cada grupo dos 652, foi selecionado o maior *contig* como representante. Identificamos que 93 desses *contigs* tinham a forma circular com tamanho de 1 a 84 kpb e os 559 restantes tinham a forma linear, com tamanho de 1 a 395 kpb. A discriminação entre *contig* circular e linear foi identificada com Platon. Apenas 17 *contigs* foram recuperados com as cinco ferramentas, 20 foram recuperados com quatro e 30 com três. Os 585 *contigs* restantes, foram recuperados apenas com duas ferramentas, principalmente com MetaPlasmidSpades e Plasflow. As ferramentas Platon e Recycler foram as mais frequentes na recuperação dos *contigs,* como mostrado na Figura 7.



Legenda: me = MetaPlasmidSpades, pw = Plasflow, pl = Platon, re = Recycler, sc = Scapp.

Figura 7: Quantidade de *contigs* de plasmídeos recuperados inicialmente por mais de uma ferramenta.

4.3.3. Busca de ORFs em *contigs* recuperados por mais de uma ferramenta

Em 3 dos 652 *contigs* recuperados por mais de uma ferramenta não foi encontrada nenhuma ORF, sendo que dois eram circulares e um era linear. Estes *contigs* sem ORF eram pequenos, com tamanho < 1,7 kpb, e foram desconsiderados nas próximas análises. Nos 649 *contigs* restantes foram encontradas 17112 ORFs e estes *contigs* foram submetidos a anotação com três bancos de dados. Obtivemos 17013 ORFs alinhadas com as sequências de referência, dos quais 8571 (50%) apresentaram similaridade contra as proteínas do DEG, 13201 (77%) contra o RefSeq e 12925 (76%) contra o COG, com e-value < 1e-6 e bitscore > 50. Desta forma, 13840 ORFs apresentaram similaridade aceitável, das quais 8509 contra os três bancos de dados (Figura 8).

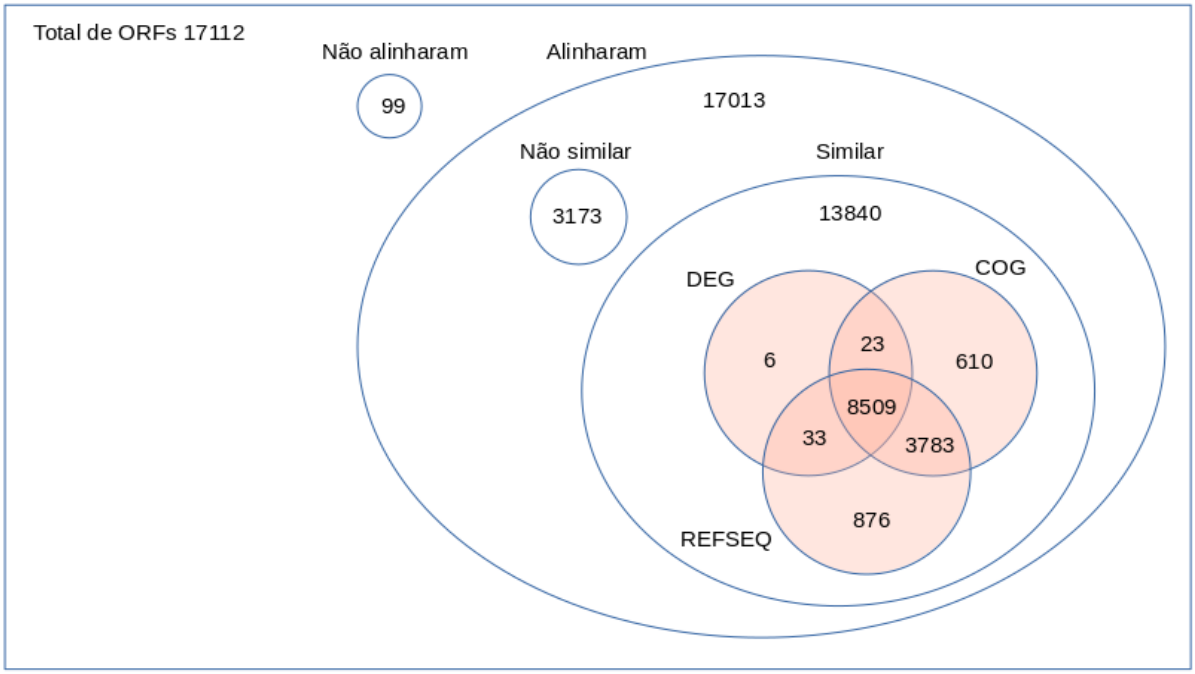


Figura 8: Diagrama de Venn dos 17112 ORFs encontrados em 649 *contigs* recuperados inicialmente Similaridade e-value < 1e-6 e bitscore > 50 contra três bancos de dados.

Observamos que a frequência de genes essenciais identificados pelo DEG nos *contigs* circulares foi menor do que nos *contigs* lineares (Figura 9). A frequência de genes essenciais por *contig* circular na maioria das vezes variou de 0 a 20%. Por outro lado, nos *contigs* lineares a frequência variou de 0 a 100% com uma maior ocorrência entre 50%-60%. Os pontos discrepantes nos *contigs* circulares, correspondem a alguns *contigs* com tamanho < 3,5 kpb com uma frequência de genes essenciais > 20%. Segundo Samuel & Sebastian (2015), genes essenciais, como aqueles codificadores de proteínas estruturais

ou relacionados a funções metabólicas básicas, raramente estão localizados em plasmídeos.

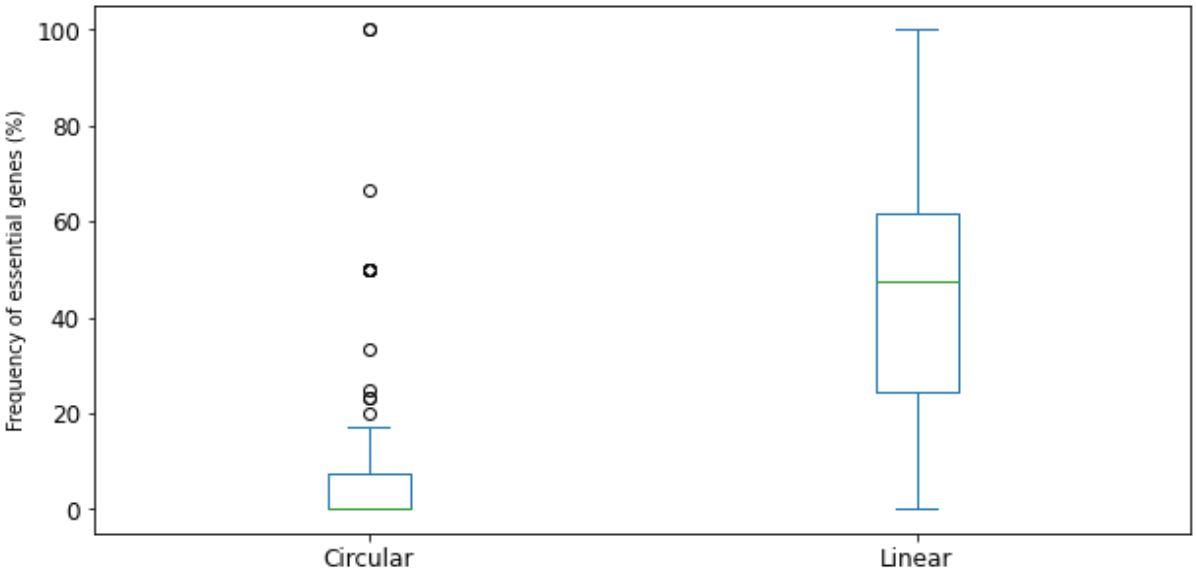


Figura 9: Comparação da frequência de genes essenciais entre *contigs* circulares e lineares.

Nos *contigs* com frequência de até 100% de genes essenciais, as categorias funcionais E, J, M, R, C e G foram as mais abundantes (em ordem decrescente), e as categorias T, H, L, K, O, P, I, S, D, V, F, N e X, aparecem com abundância intermediária. As categorias U, Q, Z, A e W foram menos abundantes. Não houve nenhum caso de genes classificados nas categorias B e Y (Figura 10). Consideramos como nível mais abundante todas as categorias com abundância de genes ≥ 800, nível intermediário com abundância entre 200 a < 800 e menos abundante < 200.

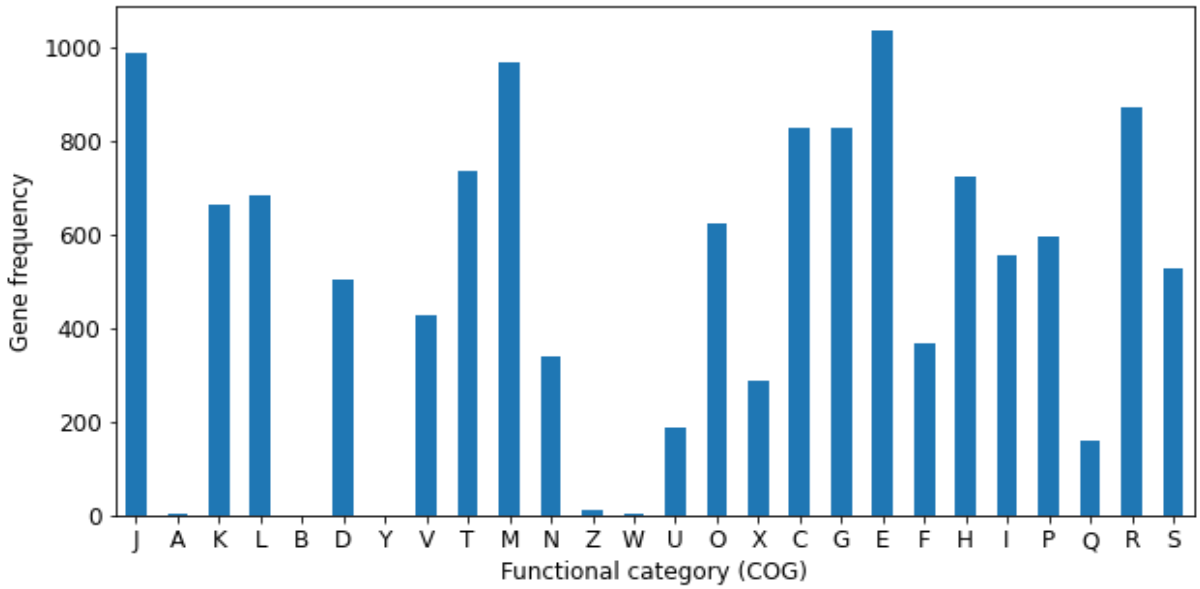


Figura 10: Distribuição da frequência de genes (número de genes) por categoria funcional (COG).
J - Tradução, estrutura ribossômica e biossíntese; A - Processamento e modificação de RNA; K - Transcrição; L - Replicação, recombinação e reparo; B - Estrutura e dinâmica da cromatina; D - Controle do ciclo celular, divisão celular, partição cromossômica; Y - Estrutura nuclear; V - Mecanismos de defesa; T - Mecanismos de transdução de sinal; M - Parede celular/membrana/biosintese do envelope; N - Motilidade celular; Z - Citoesqueleto; W - Estruturas extracelulares; U - Tráfego intracelular, secreção e transporte vesicular; O - Modificação pós-traducional, renovação de proteínas, chaperonas; X - Mobiloma: profagos, transposons; C - Produção e conversão de energia; G - Transporte e metabolismo de carboidratos; E - Transporte e metabolismo de aminoácidos; F - Transporte e metabolismo de nucleotídeos; H - Transporte e metabolismo de coenzimas; I - Transporte e metabolismo lipídico; P - Transporte e metabolismo de íons inorgânicos; Q - Biossíntese, transporte e catabolismo de metabólitos secundários; R - Função geral; S - Função desconhecida.

Suzuki *et al*. (2019), analisaram o plasmidoma do microbioma do intestino humano, e revelaram que as categorias X, S, P, V e U são mais abundantes em plasmídeos do que nos cromossomos bacterianos, e a categoria G que corresponde ao transporte e metabolismo de carboidratos é mais abundante nos cromossomos do que nos plasmídeos. Observamos que a maior parte dos *contigs* lineares que recuperamos como potenciais plasmídeos tinham uma grande quantidade de genes que codificavam para a categoria G, a qual está fortemente associada com os genes essenciais. Assim, optamos por separar os *contigs* pelo tipo de categoria funcional e darmos preferência para os *contigs* com alta frequência das categorias que possuem uma forte relação com os plasmídeos, principalmente a categoria X.

Buscamos uma forma de separar os *contigs* por categoria funcional, e verificamos que à medida que removemos os *contigs* por quantidade de genes essenciais, a frequência das categorias funcionais tende a diminuir. Porém os *contigs* que contêm os genes que codificam para a categoria X não diminuem na mesma proporção do que os *contigs* que possuem os genes para as demais categorias, como pode ser visto na Figura 11. Por

exemplo, quando removemos os *contigs* contendo 50% de genes essenciais, a frequência da categoria X é ainda de 90%, enquanto as demais categorias ficam entre 19% (categoria J) e 49% (categoria V). Observamos que a categoria J é a mais afetada praticamente em todos os casos. A categoria L tende a ser menos afetada entre 55% e 15%. Esses resultados indicam que os *contigs* que possuem a categoria X estão menos associados com os *contigs* que possuem as demais categorias e por meio da presença da categoria X é possível selecionar com mais chance de acerto os potenciais *contigs* de plasmídeos. Na Figura 11, as categorias Z, A, W, B e Y foram omitidas por terem sido pouco representadas (< 10 genes).

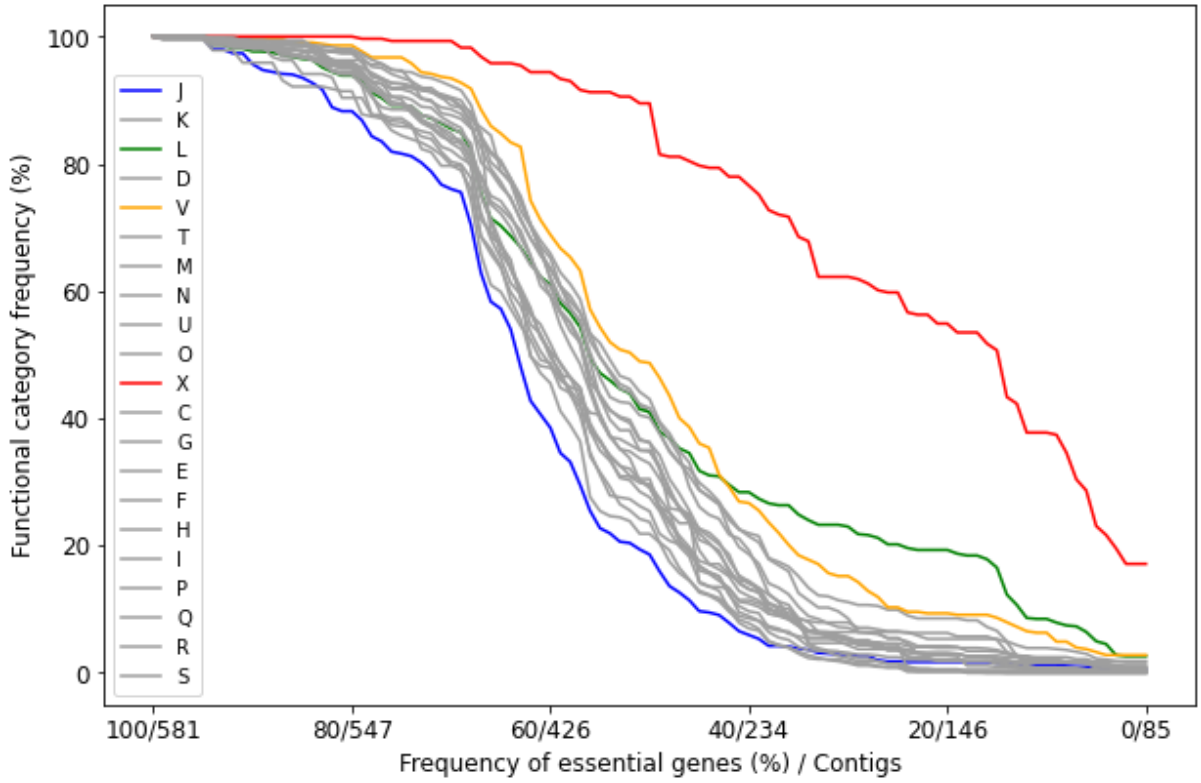


Figura 11: Diminuição da frequência da categoria funcional (eixo y) à medida que os *contigs* são removidos pela frequência (%) de genes essenciais (eixo x).

Outra observação que também nos chamou a atenção foi a distribuição das categorias pelo tamanho dos *contigs* (Figura 12). Verificamos que a categoria X está mais presente nos *contigs* menores. Por exemplo, em *contigs* com tamanho de até 5 kpb, a categoria X é representada por 70 genes, e à medida que o tamanho dos *contigs* aumenta, o número dessa categoria tende a diminuir, ao ponto que a partir do tamanho de 50 kpb, a ocorrência da categoria X é muitíssimo reduzida.

A categoria L é a segunda mais presente nos *contigs* menores que 5 kpb, no entanto, observamos que essa categoria não diminui à medida que os *contigs* aumentam de tamanho (Figura 12). A maioria das categorias estava mais presente nos *contigs* com tamanho de 10 a 70 kpb, no entanto, as categorias J, M, C e E, foram encontradas principalmente nos *contigs* maiores.

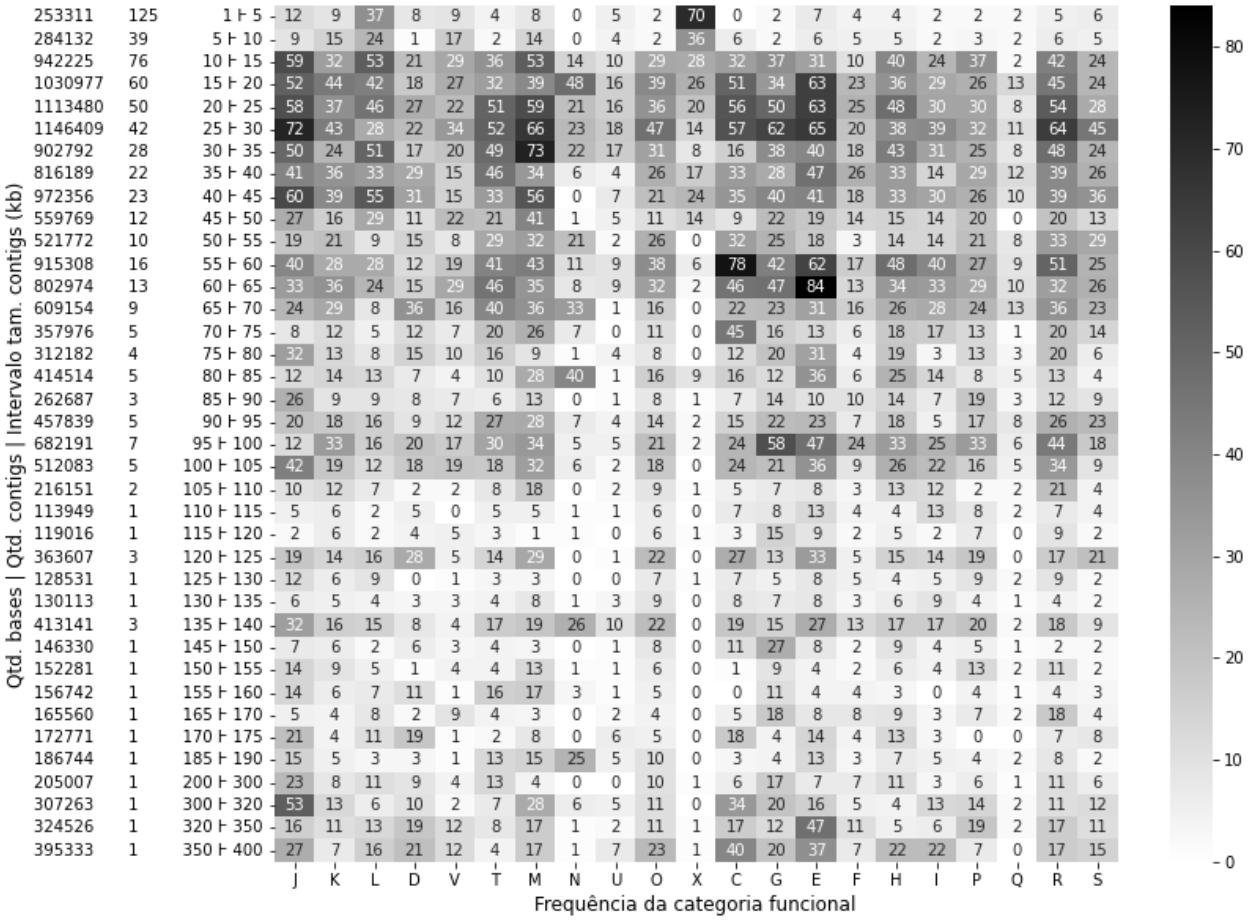


Figura 12: Distribuição da frequência de categorias funcionais (COG) por tamanho dos *contigs*.
No eixo y são mostradas 3 colunas que correspondem da esquerda para a direita: Qtd. Bases – quantidades de bases; Qtd. *Contigs* – quantidade de *contigs*; Intervalo do tamanho dos *contigs*. Vários intervalos foram omitidos pela ausência de *contigs*.

4.4. Seleção de *contigs* de plasmídeos

Levando em consideração a anotação, a frequência das categorias funcionais, a forma (circular ou linear) e tamanho dos *contigs* filtramos manualmente os *contigs* com características mais aceitáveis de plasmídeos, principalmente aqueles que continham genes responsáveis pela manutenção e replicação plasmidial como os genes parA, parB, repA, repB, repL e repM que já estão bem descritos na literatura (Nazir *et al*., 2020),

relaxases como das famílias MobA e MobL (Monzingo *et al*., 2007), proteínas do sistemas de secreção tipo IV (T4SS) e proteínas de acoplamento tipo IV (T4CP) (Wallden *et al*, 2010). A distribuição dos genes por categoria funcional (COG) no conjunto de *contigs* filtrados (230 *contigs* listados no Anexo 1) está mostrada na Figura 13.

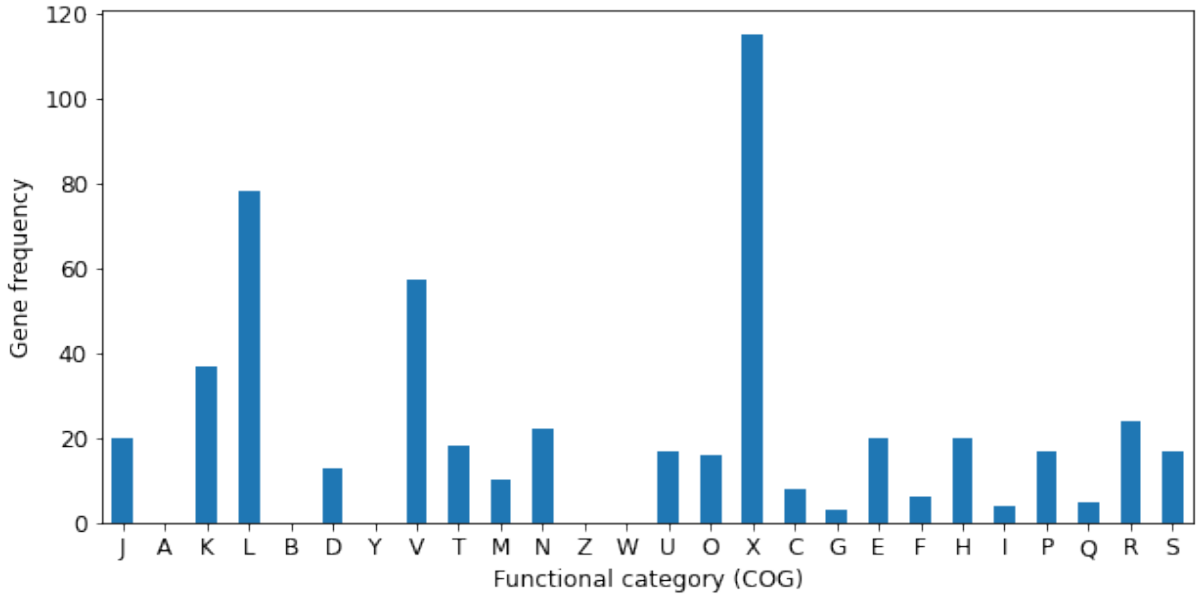


Figura 13: Distribuição da frequência de genes (número de genes) por categoria funcional (COG) após a adoção de um filtro manual dos *contigs* característicos de plasmídeos.

Após o filtro manual adotado para a seleção dos *contigs* de plasmídeos, sem que a proporção de 40% dos genes na categoria X fosse alterada, observamos uma redução de 287 para 115 genes classificados nesta categoria (Figuras 10 e 13). A redução deve-se ao fato de que 24 *contigs* foram removidos por corresponderem a sequência de bacteriófagos (codificam proteínas estruturais como capsídeo e cauda de vírus, além de outras proteínas como replicases e terminases).

Assim como ocorreu com a categoria X, após o filtro, a frequência das categorias K, L, V, N, U, R e S foram menos afetadas. Por outro lado, as categorias J, M, C, G, E terminasse F foram mais afetadas em comparação com os valores iniciais mostrados na Figura 10. Após a seleção manual, obtivemos 230 *contigs* sendo 78 circulares com tamanho entre 1 kpb a 13,6 kpb e 152 *contigs* lineares com tamanho entre 1 a 19,6 kpb (Anexo 1).

4.4.1. Anotação dos *contigs* de plasmídeos

A anotação dos 230 *contigs* de plasmídeos revelou que 47% das proteínas anotadas correspondem a proteínas hipotéticas e proteínas sem similaridade, isso mostra que ainda

existe uma grande fração de proteínas desconhecidas que podem ser futuramente exploradas. Das proteínas com similaridade nos bancos de dados, as transposases da categoria X, proteínas de partição e início de replicação plasmidial, recombinação e reparo de DNA da categoria L e proteínas relacionadas com o mecanismo de defesa da categoria V foram as mais abundantes. A alta abundância de genes de transposases em *contigs* com tamanho de até 5 kpb pode explicar o grande índice de fragmentação dos *contigs* na etapa de montagem já que são sequências flanqueadas por repetições que dificultam a montagem de *contigs* longos (Claire *et al*., 2022). Na Tabela 6 estão as principais anotações que foram realizadas.

Tabela 6: Anotação dos *contigs* de plasmídeos.

Gene	Atividade	Função
srtR	XRE regulação transcricional	Estresse oxidativo e fator de virulência
*	Transposase	Transposição de DNA
YafQ/VapC/RelE/ParE/RelB/DinJ/RatA	Toxina-antitoxina	Manutenção plasmidial
*	Tipo P ATPase	Translocação de cadmio
YgaP	Tiossulfato sulfurtransferase	Destoxificação de cianeto
TerB	Telurito resistente	Resistência antimicrobiana
XcpX	T2SS	Pseudo pilina
MobA/MobL	Relaxase	Conjugação plasmidial
YqgE/AlgH	Regulação transcricional	Fator de virulência
YitT/sigB	Regulação transcricional	Tolerância a estresse
WYL domínio	Regulação transcricional	Estresse e reparo do DNA
CopG	Regulação transcricional	Controle de número do cópias de plasmídeos
LexA	Regulação transcricional	Resposta SOS (reparo de DNA)
SoxR	Regulação transcricional	Redox
PadR	regulação transcricional	Destoxificação de ácidos fenólicos
ArsR/SmtB	Regulação transcricional	Resposta ao estresse em decorrência a metais pesados
MerR	Regulação transcricional	Resistência ao mercúrio
MarR	Regulação transcricional	Estresse e virulência
LysR	Regulação transcricional	Virulência e *quorum sensing*
IclR	Regulação transcricional	Degradação de compostos aromáticos
GntR	Regulação transcricional	Patogenicidade e fator de virulência
ArsR	Regulação transcricional	Resposta ao estresse metálico
TonB	Receptor externo da membrana	Transporte de compostos aromáticos
TauE/SafE	Proteína de membrana	Efluxo de sulfito (destoxificação)
PrsW	proteína de membrana	Resposta ao estresse
LysE	Proteína de membrana	Transporte de metais pesados
repL/repM/repB/repA/parA/parB	Proteínas de ligação ao DNA	Partição e manutenção plasmidial
Lnu(P)	Nucleotidiltransferase	Resistência a lincosamida

NERD	Nuclease pXO1 de Bacillus anthracis	Fator de virulência
PilB/PilD/PilQ/PilV/PilT/pilO/pilH/ pilG/pilM	Pilus tipo IV	Motilidade e reconhecimento de célula receptora
UvrD	Helicase	Reparo de DNA sob Luz UV
YheC/YheD	Esporulação	Mecanismo de defesa
*	DNA metiltransferase	Metilação de DNA
LanM (type II)	Biosintese de lantipeptide	Antimicrobiano
EgtB	Biosintese de ergotioneína	Sulfurização oxidativa
AS-48ª	Biosintese de bacteriocina	Antimicrobiano
HicB (type II)	Antitoxina	Mecanismo de defesa
*	CRISPR-cas	Mecanismo de defesa
*		ATPase AAA
*		Integrase
IstB		IS21-like element helper ATPase
*		Manganês-catalase
MBL		Metalohidrolase
M48		Metaloprotease
RuvX		Resolvase
Trbl (type F)	dissulfeto isomerase	Conjugação plasmidial
TssG/TssF	T6SS	Sistema de secreção tipo VI
RnfH		Fixação de nitrogênio
*		Recombinação de DNA
*		fago antirepressor
*		Redução de nitrito
*		Oxidação de cobre
*		Transporte de magnésio
GspH/FimT		Pseudo pilina
RecN/ RadC		Reparo de DNA
DinB		Destoxificação dependente de tiol
*		Efluxo de arsênio
APH(6)/APH(3)		Resistência a aminoglicosídeos
Sul1		Resistência a sulfonamida

* genes (ORFs) sem nomes

Genes codificadores de proteínas de regulação transcricional associados a diferentes formas de estresse, fatores de virulência, reparo de DNA, destoxificação de compostos como ácidos fenólicos, transporte e degradação de composto aromáticos e resistência contra metais pesados como mercúrio foram abundantes. Por outro lado, em menores quantidade foram encontrados genes de proteínas de oxidação de cobre, fixação de nitrogênio, redução de nitrito, transporte de magnésio, antirepressor de fago, enzimas de restrição, integrase, resolvase, metaloprotease, manganês-catalase, e resistência a sulfonamida.

Em quantidades intermediarias estavam, as proteínas YafQ, VapC, RelE, RelB, ParE, DinJ YoeB, e RatA do sistema toxina-antitoxina do tipo II. Esse sistema foi associado à manutenção plasmidial e já foram reportados em plasmídeos de águas subterrâneas por

Kothari *et al*., (2019) e em ambientes extremos por Perez *et al*., (2021). YoeB e RelE são toxinas do tipo RNase dependentes de ribossomos que se ligam diretamente ao sítio A do ribossomo, onde clivam o mRNA associado ao ribossomo (Page & Peti, 2016). Com alta frequência foram identificados os genes repL, repM, repB e repA responsáveis pela replicação plasmidial e os genes parA e parB reportados em plasmídeos de baixo número de cópias que atuam no mecanismo de partição como nos plasmídeos P1, P7, p42d, PMOL28 e PRiA4b (Bignell & Thomas, 2001).

Proteínas associadas à resistência contra o cádmio, arsênio e telurito foram identificadas e foram anotadas como proteínas de membrana, principalmente aquelas que estão integradas a bombas de efluxo. A destoxificação de cianeto e sulfito pode ocorrer pela atividade de tiossulfato sulfurtransferase e bombas de efluxo, respectivamente, e tais componentes foram anotados nos *contigs* de plasmídeos. Também foram anotados genes de proteínas relacionadas com o sistema de defesa CRISPR-cas1-8, metilação do DNA, esporulação, reparo de DNA induzido por luz UV, resposta ao estresse oxidativo, resistência enzimática de aminoglicosídeos e lincosamida, assim como proteínas de biossíntese de antimicrobianos como lantipeptideo e bacteriocinas. Foram identificadas proteínas pertencentes ao sistema de secreção tipo VI (T6SS) e tipo II (T2SS), genes PilB, PilD, PilQ, PilV, PilT, pilO, pilH, pilG e pilM de proteínas que formam o *pilus* tipo IV (T4P, Type IV Pilus). Esse *pilus* está associado com diferentes comportamentos da bactéria, incluindo motilidade, formação de biofilme, secreção de proteínas e quimiotaxia, fator de virulência e reconhecimento da célula receptora (Craig *et al*., 2019). Proteína relaxase do tipo MobA/MobL e dissulfeto isomerase codificado pelo gene Trbl encontrado em plasmídeos, também foram anotadas. O gene Trbl foi relatado por Maneewannakul *et al*. (1992), e faz parte do sistema de secreção tipo IV (T4SS) que é responsável pela conjugação plasmidial. Também foram anotadas proteínas que participam da biossíntese e transporte pela membrana de ácido fólico (vitamina B12) assim como também aquelas envolvidas no metabolismo de compostos aromáticos como fenóis. Outras proteínas que não são mostradas na Tabela 6 também foram identificadas como a *thiol-disulfide oxidoreductase* (envolvida na atividade antioxidante), enzimas de degradação de ácidos graxos de cadeia longa, produto proteico da ORF-3, do gene virA encontrado normalmente no plasmídeo pRiA4 da bactéria *Agrobacterium rhizogenes*, responsável pela tumorigênese em plantas (Endoh *et al*., 1990).

4.4.2. *Contigs* de plasmídeos recuperados na forma circular

Como já mencionado, recuperamos 78 *contigs* de plasmídeos circulares. A seguir descrevemos alguns desses *contigs* e os genes que foram anotados.

Os *contigs* p617 [NODE_617_length_5550_cov_11.193067], encontrado na amostra do dia 01, e p1234 [NODE_1234_length_5548] da amostra do dia 64, ambos com 5,5 kpb são bastante similares (99% de identidade e 100% de cobertura). A diferença entre eles é de apenas duas bases, pois p1234 não tem duas adeninas, uma na posição 1706 e a outra na posição 4454. A forma circular e a presença do gene parA indicam fortemente que esses *contigs* são de origem plasmidial. Diante da alta similaridade entre as sequências, consideramos que eles correspondem ao mesmo plasmídeo (um desses *contigs* não foi excluído na etapa inicial que removeu *contigs* 100% idênticos e manteve o representante maior). Esse plasmídeo, ainda não foi reportado na base de dados do NCBI.

A Figura 14 apresenta as características do p617, incluindo as repetições do plasmídeo localizadas em regiões opostas, numa distância de 1022 bases uma da outra. Fora das regiões repetidas, foram encontrados dois genes repA diferentes com tamanho de 735 e 879 bases respectivamente. Este plasmídeo contém o gene MerR (associado com a resistência ao mercúrio e regulação transcricional) que também foi encontrado em *contigs* lineares. No p617 este gene está integrado entre repetições duplas (Figura 15).

No plasmídeo p617 também foram encontrados cinco sítios de restrição para as enzimas BamHI, EcoRI e PstI (Figura 16). O sítio EcoRI está localizado dentro de um dos genes parA e o sítio PstI está localizado dentro do outro gene parA. Também foram encontradas quatro repetições em tandem triplamente repetidas com tamanho de 9 bases, sendo que três dessas repetições correspondiam a sequência TTGTTGTTG, e uma correspondia a sequência CGCCGCCGC. Pelo tamanho e características das repetições, supomos que se trata de sequências que correspondem a microssatélites. Houve um caso no qual as repetições CGC e TTC foram localizadas respectivamente nas extremidades de um dos genes repA.

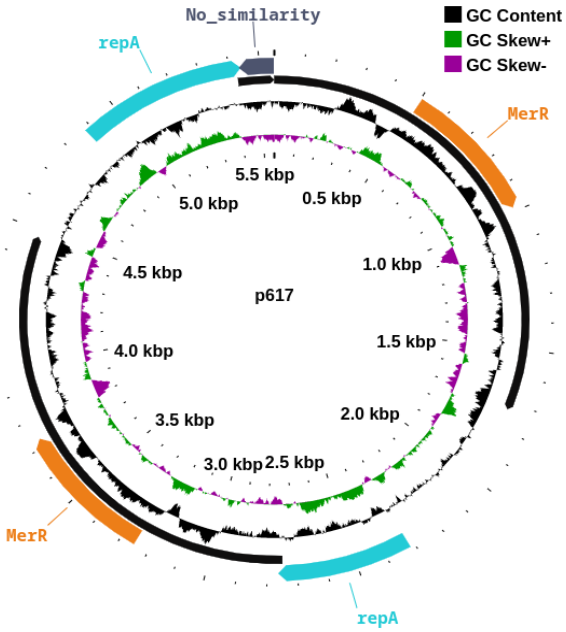


Figura 14: Plasmídeo p617 circular, encontrado no dia 01 contendo repetições em regiões opostas (segmento em preto) integrando genes de resistência ao mercúrio (seta laranja).

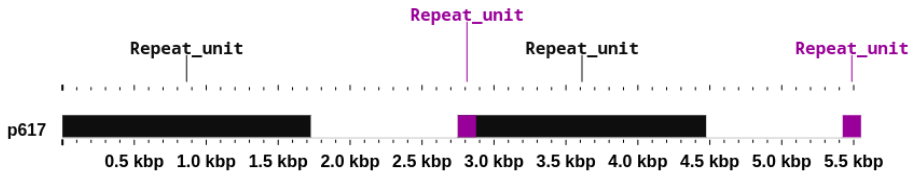


Figura 15: *Contig* p617 mostrado em forma linear com repetições que contém o gene MerR

Na amostra do dia 15, recuperamos um *contig* circular com tamanho de 2,8 kpb. Esse *contig* tem um gene repA e um gene MerR ambos integrados numa sequência bastante similar à região dos plasmídeos p617 e p1234, inclusive, quanto aos sítios de restrição BamHI, EcoRI e a repetição em tandem TTGTTGTTG, além de duas sequências diretamente repetidas de 127 bases com 100% de identidade, uma localizada no início do *contig* e a outra no final. Esse *contig* também não apresentou nenhuma similaridade com o banco de dados do NCBI. Por tanto se trata de um plasmídeo desconhecido e o chamamos de p2134 (Figura 16), sendo que pode ter alguma relação com o plasmídeo p617.

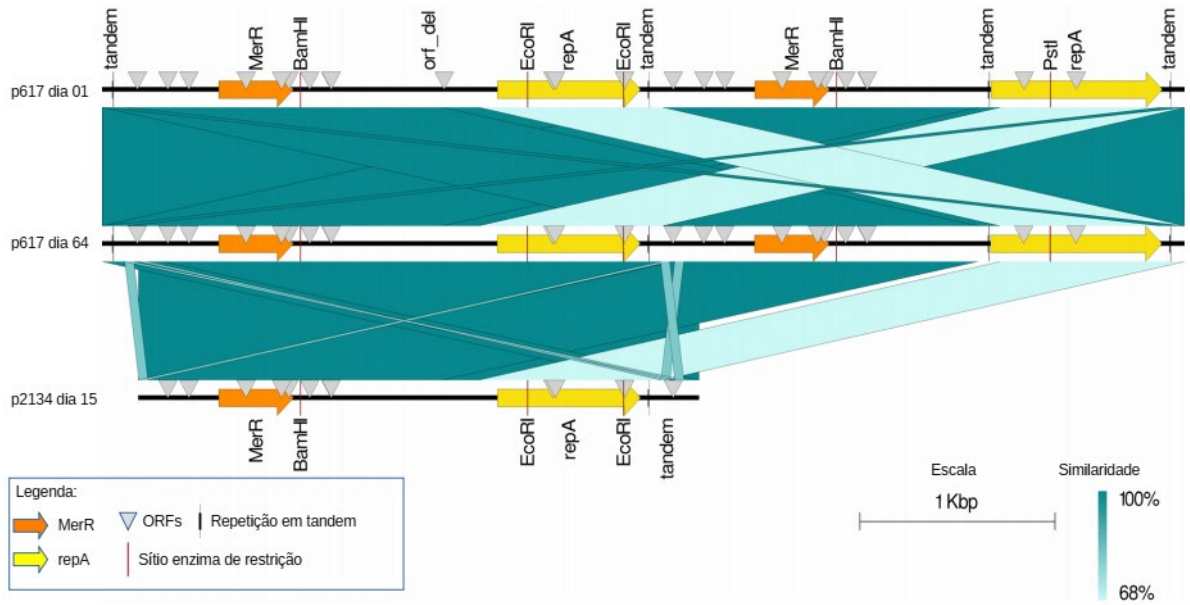


Figura 16: Comparação entre o plasmídeo p617 encontrado no dia 01 e 64 e o plasmídeo p2134 encontrado no dia 15.

Outros 3 plasmídeos circulares (p180, p1522 e p611) que recuperamos estão esquematizados na Figura 17. No plasmídeo circular p180 de 7 kpb foram anotados os genes MerR de resistência ao mercúrio, gene de recombinação de DNA e DndB associado com a proteção do DNA a estresses ambientais como temperatura, salinidade, pressão, UV, raio X, metais pesados e pH (Yang *et al*., 2017). No plasmídeo p1522 de 3,4 kpb, observamos a ocorrência simultânea do gene repA e de um gene que codifica uma integrase, indicando que esse plasmídeo pode adquirir novos genes com diferentes funções. No plasmídeo circular p611 de 11 kpb, os genes ArsR e SmtB responsáveis pela regulação transcricional, conferem resistência a metais pesados, genes TauE e SafE que participam no efluxo de sulfito, o gene Ygap associado com a destoxificação de cianeto, proteína MBL metalo-hidrolase (que pode estar associada com a resistência a beta-lactâmicos), proteína nitrito redutase e transposase que pode estar facilitando a disseminação desses genes para outros plasmídeos

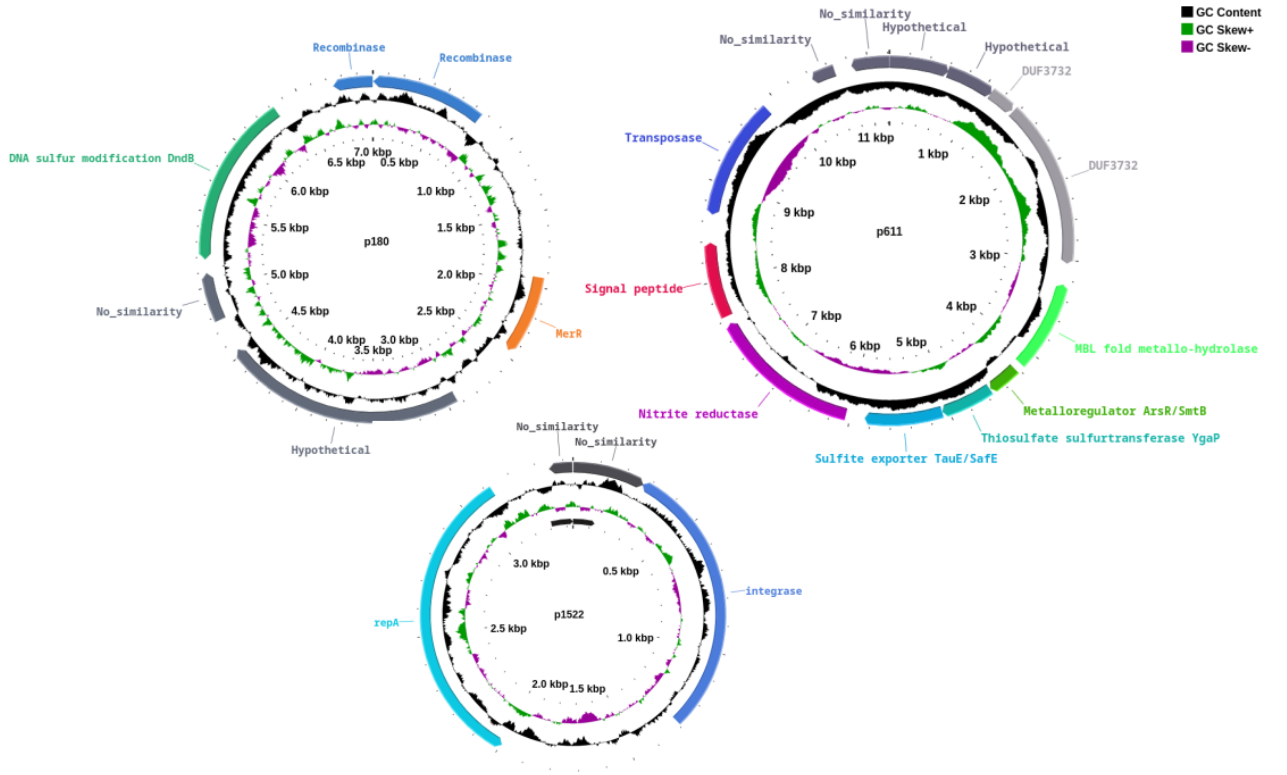


Figura 17: Plasmídeos ainda desconhecidos carregando genes acessórios que podem favorecer a célula hospedeira sobre estresse ambiental.

4.4.3. *Contigs* de plasmídeos recuperados na forma linear

Alguns *contigs* de plasmídeos que foram recuperados na forma linear estão esquematizados na Figura 18. Nos *contigs* lineares p491 de 12 kpb e p606 de 6 kpb foram anotados genes de Lantibióticos tipo II que possuem atividade antimicrobiana (Huan & Wilfred, 2012) e os genes AbiEii e AbiGii do sistema de toxina tipo IV que protege as bactérias da propagação quando são infectadas por bacteriófagos (Dy *et al*., 2014). Também foram anotados genes de transposases. Os genes de relaxases, que consistem em proteínas responsáveis pela mobilização do DNA, toxina-antitoxina tipo II, proteínas de transporte de cianeto, efluxo de metais pesados e redox transcricional foram anotados nos *contigs* p691 e p700, ambos com cerca de 9 kpb. A presença de genes de proteínas de mobilização, partição e início de replicação plasmidial é um forte indicativo que esses *contigs* são de plasmídeos. Em outros *contigs* lineares menores foram anotados genes de resistência a metais pesados, como arsênio e mercúrio, resistência contra aminoglicosídeos e sulfonamidas, recombinases, proteínas de partição e conjugação plasmidial e stress oxidativo.

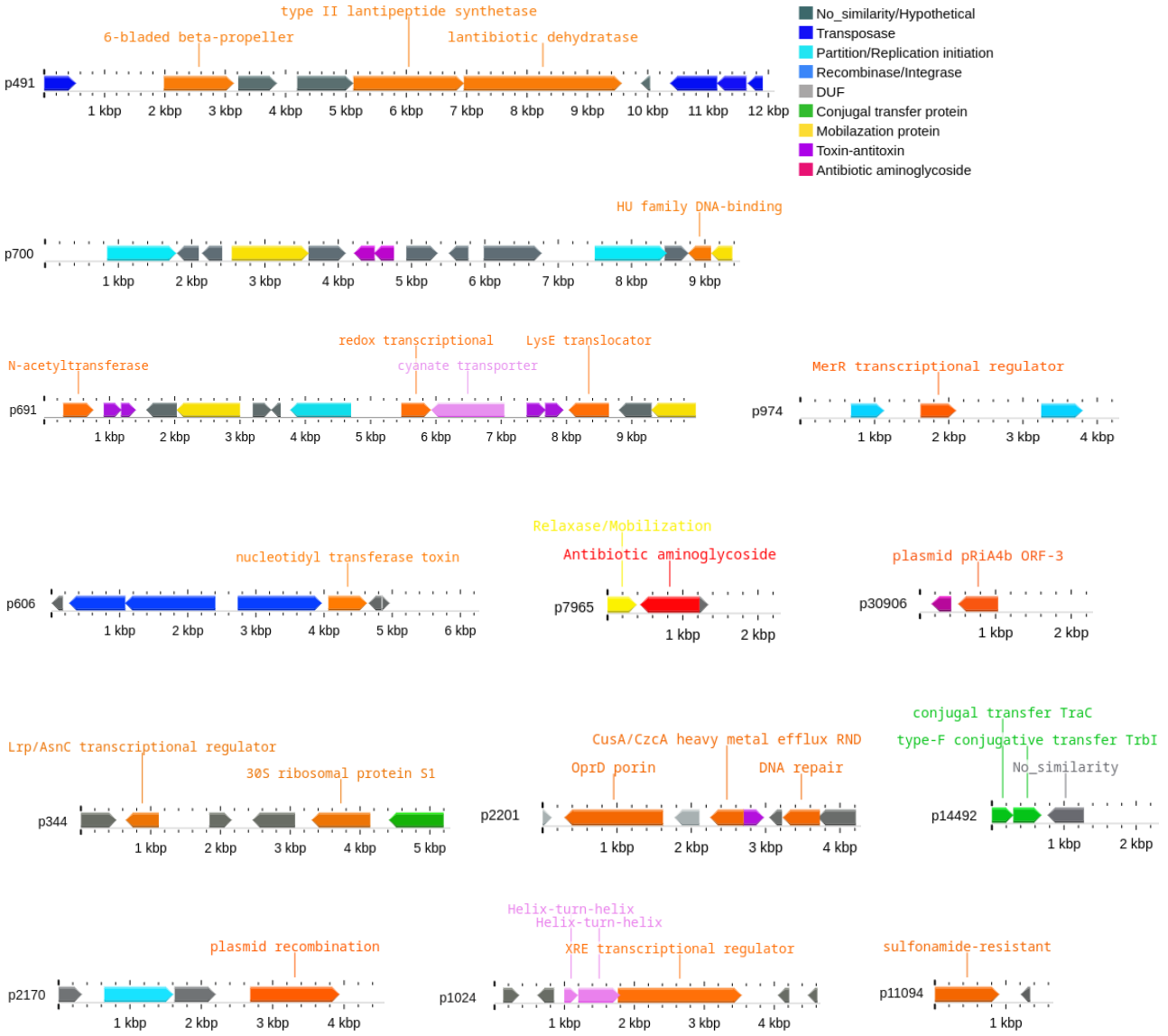


Figura 18: Anotação de *contigs* de alguns plasmídeos recuperados na forma linear.

4.4.4. Identificação de plasmídeos conhecidos nos *contigs* de plasmídeos

As análises para verificarmos a ocorrência de plasmídeos já conhecidos no conjunto final de 230 *contigs* recuperados dos metagenomas da compostagem ZC4 foram realizadas com o Nucmer contra a base de dados de plasmídeos do RefSeq que contava com 22796 sequências de nucleotídeos com tamanho de 1 kpb a 500 kpb. Sequências da base de dados com tamanho < 1 kpb e > 500 kpb foram desconsideradas, porque não permitem uma classificação confiável e raramente correspondem a plasmídeos (Schwengers *et al*., 2020). Nesta análise foram recuperados 32 *contigs* lineares com tamanho de até 2,7 kpb e identidade ≥ 90% contra regiões curtas de 2392 plasmídeos da base de dados com tamanho de 2,7 a 499 kpb. A maior cobertura observada foi de 50,9% que ocorreu com um

plasmídeo de 2,7 kpb da base de dados. Observamos que um mesmo *contig* tinha alta similaridade contra várias regiões de um mesmo plasmídeo da base de dados, assim como também tinha alta similaridade a regiões de outros plasmídeos, sendo que tais *contigs* correspondiam a genes frequentes em plasmídeos. Uma minoria de *contigs* tiveram alinhamentos únicos. Na maior parte estes *contigs* codificavam proteínas hipotéticas, transposases, proteínas de início de replicação de plasmídeos, resistência a antibióticos e recombinases. Em uma menor proporção codificavam proteínas de conjugação, mobilização plasmidial e sistemas de toxina-antitoxinas.

Com outra estratégia para pesquisar a existência de plasmídeos conhecidos nos metagenomas da compostagem, fizemos o alinhamento das *reads* de cada uma das amostras de ZC4 contra a base de dados de plasmídeos do RefSeq, utilizando a ferramenta Bowtie2. Nesta análise, apenas 11 plasmídeos da base de dados tiveram cobertura de mapeamento ≥ 90%, com 52 a 970 *reads* alinhados por plasmídeo (Tabela 7). A maioria desses plasmídeos tinha tamanho entre 2,4 a 4,3 kpb, sendo que apenas o plasmídeo NZ_CP048309.1 tem tamanho maior (24,5 kpb). Os plasmídeos NZ_LN873256.1 e NZ_LN873255.1 diferem apenas em um par de bases e carregam um gene de resistência a aminoglicosídeos.

Alguns plasmídeos foram detectados em mais de uma amostra da série temporal da compostagem. Por exemplo, o plasmídeo NZ_CP050431.1 está presente em 8 amostras da compostagem, exceto na amostra 67. O plasmídeo NC_022993.1 está presente em 3 amostras. A maioria desses plasmídeos são circulares e depositados como sequências completas no RefSeq. Apenas o plasmídeo NZ_CP035326.1 tem estrutura linear. O plasmídeo NZ_CP048309.1 carrega um gene de resistência aos beta-lactâmicos.

Dos 11 plasmídeos mapeados com Bowtie2, 4 também foram encontrados na análise com Nucmer. Nos casos em que não foram encontrados *contigs* associados com os plasmídeos mapeados, uma possível explicação é que os *contigs* para esses plasmídeos tinham tamanho < 1 kpb e foram excluídos da análise. Dessa forma, concluímos que Bowtie2 (que usa *reads*) mostrou ser mais adequado na busca por plasmídeos já conhecidos do que o Nucmer (que usa *contigs*).

Tabela 7: Plasmídeos conhecidos encontrados pelo mapeamento das *reads* com Bowtie2.

Plasmídeos no RefSeq	Forma	Tamanho.	*Reads*	Amostra do Dia
* NZ_LN873256.1 *Acinetobacter lwoffii* strain ED23 35 plasmid pALWED1.8	circular	4135	540	64
* NZ_LN873255.1 *Acinetobacter johnsonii* strain LS47 1 plasmid pAJOLS1.1	circular	4136	526	64
* NZ_LN873255.1 *Acinetobacter johnsonii* strain LS47 1 plasmid pAJOLS1.1	circular	4136	102	67
NZ_CP050431.1 *Acinetobacter baumannii* PM194188 plasmid pPM194122_6	circular	2419	737	01
NZ_CP050431.1 *Acinetobacter baumannii* PM194188 plasmid pPM194122_6	circular	2419	603	03
NZ_CP050431.1 *Acinetobacter baumannii* PM194188 plasmid pPM194122_6	circular	2419	358	07
NZ_CP050431.1 *Acinetobacter baumannii* PM194188 plasmid pPM194122_6	circular	2419	642	15
NZ_CP050431.1 *Acinetobacter baumannii* PM194188 plasmid pPM194122_6	circular	2419	625	30
NZ_CP050431.1 *Acinetobacter baumannii* PM194188 plasmid pPM194122_6	circular	2419	489	64
NZ_CP050431.1 *Acinetobacter baumannii* PM194188 plasmid pPM194122_6	circular	2419	970	78
NZ_CP050431.1 *Acinetobacter baumannii* PM194188 plasmid pPM194122_6	circular	2419	879	99
NZ_CP048309.1 *Escherichia coli* strain 9 plasmid p009_E	circular	24518	752	01
* NZ_CP042565.1 *Acinetobacter baumannii* strain E47 plasmid pE47_009	circular	2427	238	64
NZ_CP035326.1 *Escherichia coli* strain BR12 DEC plasmid unnamed1	linear	3019	238	01
NZ_CP035326.1 *Escherichia coli* strain BR12 DEC plasmid unnamed1	linear	3019	52	03
NZ_CP032129.1 *Acinetobacter chinensis* WCHAc010005 plasmid p3_010005	circular	4353	132	64
NZ_CP019571.1 *Clostridium perfringens* strain CP15 plasmid pCP15_4	circular	2421	67	64
* NC_022993.1 *Geobacillus* sp. 610 plasmid pGTD7	circular	3279	166	07
* NC_022993.1 *Geobacillus* sp. 610 plasmid pGTD7	circular	3279	463	15
* NC_022993.1 *Geobacillus* sp. 610 plasmid pGTD7	circular	3279	237	64
NC_019049.1 *Escherichia coli* plasmid pCM959	circular	4012	102	01
NC_004341.2 *Weissella cibaria* plasmid pKLCB	circular	3353	179	64

*Plasmídeos encontrados também com Nucmer no conjunto de 230 *contigs* de plasmídeos recuperados (Anexo 2).

4.5. Análise de *contigs* de plasmídeos ao longo da compostagem ZC4

Os 230 *contigs* de plasmídeos foram usados como referência para o mapeamento das *reads* do metagenoma *shotgun* de cada uma das amostras da série temporal da compostagem ZC4. Observamos que 40 *contigs* estão presentes em mais de uma amostra e foram mapeados com 100% de cobertura e que 3 *contigs* estão presentes em 7 amostras (Figura 19). A maioria dos *contigs* repetidos em mais de uma amostra são pequenos (< 5 kpb). Os *contigs* > 5 kpb foram encontrados principalmente em apenas duas amostras, como ocorre com os *contigs* NODE_100, NODE_359 e NODE_444 de 16 kpb, 13 kpb e 11 kpb, respectivamente. Esse resultado mostra que *reads* de diferentes amostras podem mapear significativamente contra os *contigs* de plasmídeos montados a partir de uma determinada amostra, o que é indicativo da presença desse potencial plasmídeo ao longo da série temporal da compostagem.

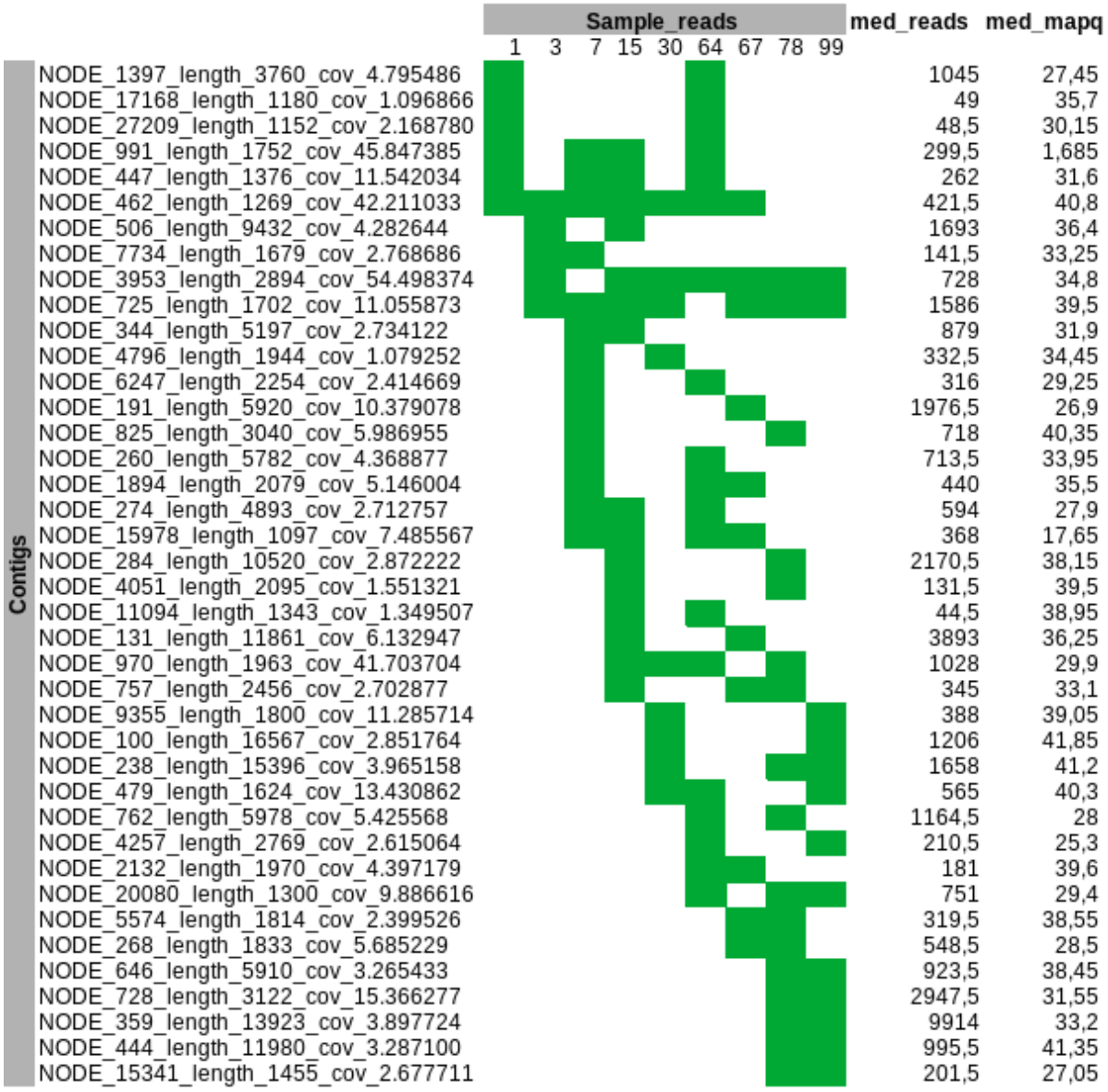


Figura 19: *Contigs* de plasmídeos que sucedem nas amostras em serie temporal da compostagem. O mapeamento de *reads* do metagenoma *shotgun* cada amostra contra os *contigs* de plasmídeos foi feito com Bowtie2 (100% de cobertura de mapeamento). med_*reads* = mediana do número de *reads*; med_mapq = mediana da média da qualidade de mapeamento. A detecção do *contig* na amostra (dias da compostagem) está destacado em verde.

4.6. Análise de genes expressos em *contigs* de plasmídeos

As *reads* do metatranscritoma foram mapeadas contra as ORFs dos 230 *contigs* de plasmídeos recuperados e o RPKM (*Reads Per Kilobase Million*) para cada ORF foi calculado. Em 82 *contigs* obtivemos mapeamento em 511 ORFs com 90% de cobertura (Figura 20) que representam 221 proteínas diferentes.

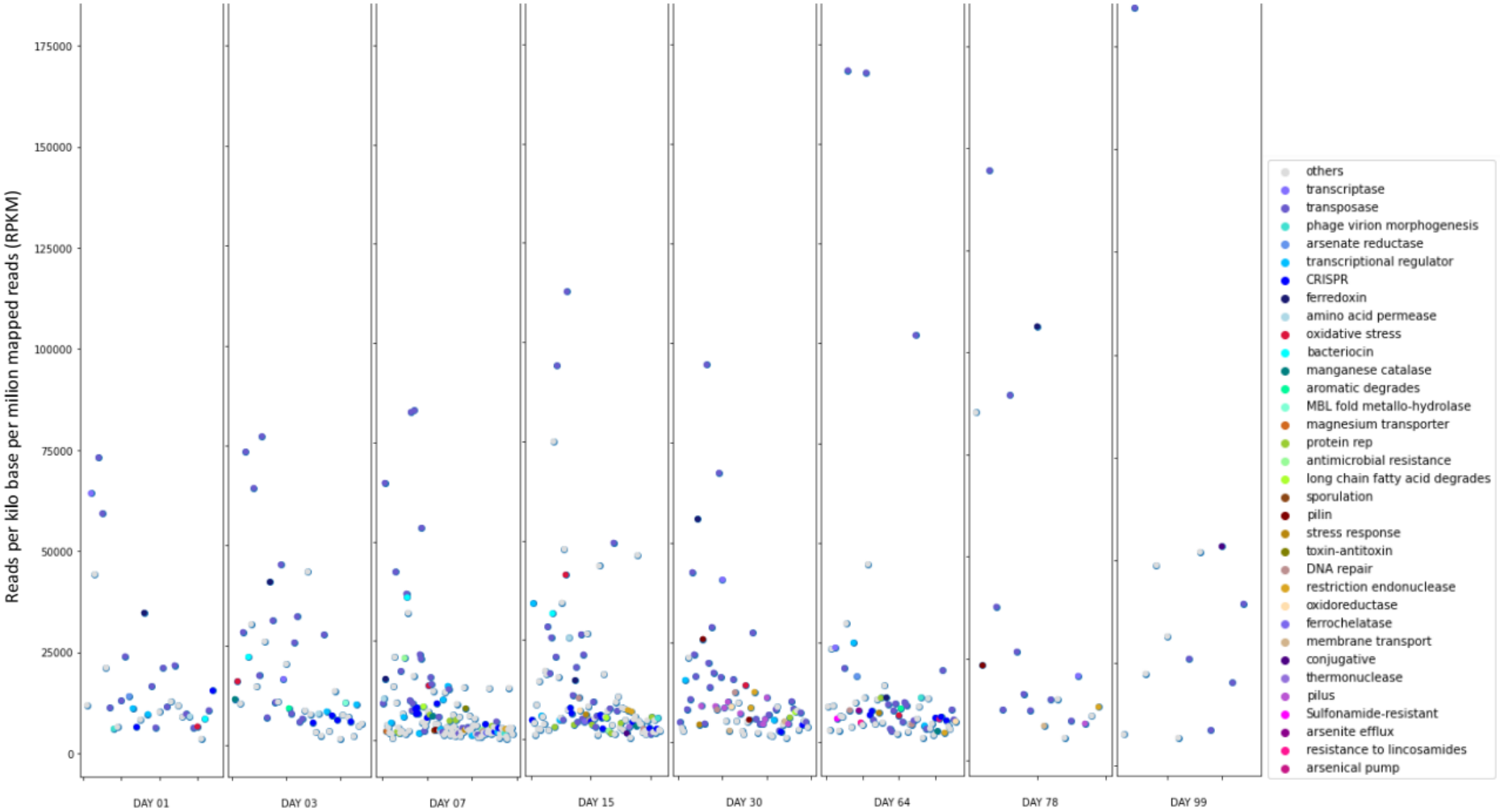


Figura 20: Avaliação de ORFs plasmidiais expressas na compostagem.
Cada ponto representa uma ORF na qual *reads* do metatranscritoma de cada amostra foram mapeados.

A maioria das ORFs teve um RPKM < 25000 em todas as amostras. As que tiveram o RPKM ≥ 25000, foram identificados como transposases associadas com as IS26, ISL3, IS1182, IS256, IS481, integrase, recombinase, proteínas de conjugação plasmidial tipo F, trbC e virB2, proteínas associadas com a destoxificação, produção de bacteriocinas e ferrodoxina.

As ORFs com RPKM menores foram identificadas associadas com arsenato redutase, esporulação, proteínas de pilus tipo 4, CRISPR, oxido redutase, manganase, catalase, proteínas do sistema toxina antitoxina, efluxo de arsenito, resistência a antibiótico principalmente a sulfonamida, proteínas de início de replicação e partição plasmidial, resistência a mercúrio e proteínas integrantes a sistemas de transporte pela membrana possibilitando o efluxo de compostos tóxicos.

4.7. Busca de ISs e Transposons

Fizemos a busca de IS e de transposons em *contigs* de plasmídeos e de cromossomos. Encontramos 5267 ISs em *contigs* de cromossomos e 74 em *contigs* de plasmídeos. Estas ISs correspondem a 209 ISs distintas pertencentes a 26 famílias. Destas 18 pertencem tanto aos *contigs* de plasmídeos como de cromossomos (Figura 21).

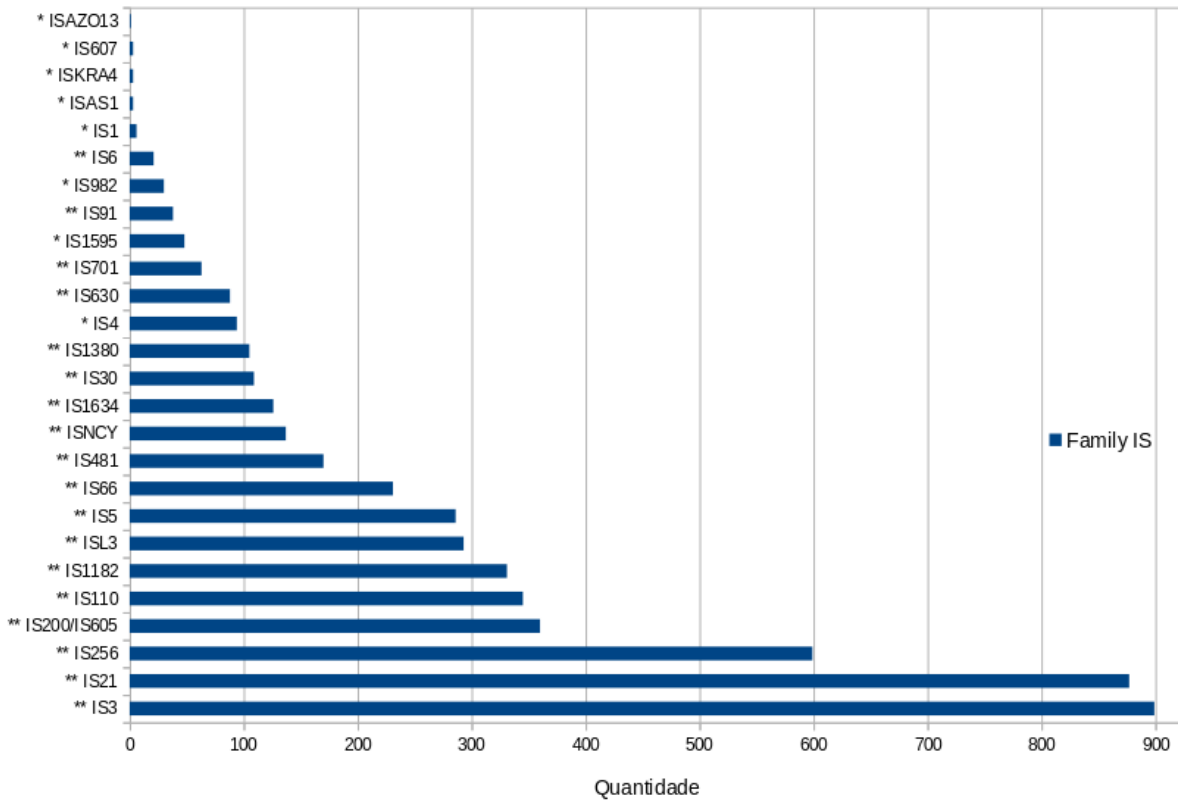


Figura 21: IS em *contigs* de plasmídeos e/ou de cromossomos.
(**) – Famílias de IS em *contigs* de plasmídeos e de cromossomos. (*) – Famílias de IS somente em *contigs* de cromossomos.

As famílias IS3, IS21 e IS256 são as mais abundantes e estão presentes tanto em *contigs* de plasmídeos como em *contigs* de cromossomos. Por outro lado, as famílias ISAZO13, IS607, ISKRA4 e IS1 são menos abundantes e estão presentes somente em *contigs* de cromossomos.

Em *contigs* de plasmídeos, encontramos várias ISs da família IS3 próximas a genes de multicobre oxidase, de LysR (regulação transcricional) e de lantibiótico *dehydratase*. A IS364 da família IS5 foi identificada próxima a genes da família ANT(3") que confere resistência aos aminoglicosídeos. A IS195 da família IS256 foi encontrada próxima a genes do sistema toxina antitoxina do tipo II, MBL *fold metallo hydrolase*, nitrito redutase, desintoxicação de cianeto, resistência ao arsênio por regulação transcricional dos genes ArsR e SmtB, exportação de sulfito e translocação de cádmio.

Nos *contigs* de cromossomos, encontramos apenas a IS176 da família IS3 próxima a um gene da família ANT(3"), confirmando que há poucos casos de ISs associadas com ARGs nos *contigs*. Dos *contigs* que carregam ISs, 97% têm tamanho < 5 kpb, mas há *contigs* com até 52 kpb. Em nenhum deles identificamos ARGs.

ISs das famílias IS21 e IS3, foram encontradas repetidas em até 3 vezes nos mesmos *contigs* com tamanho de até 5,6 kpb. Segundo Babakhani & Oloomi (2018), um transposon composto é constituído pelo menos por duas ISs do mesmo tipo, normalmente flanqueiam um ou mais genes, possuem tamanho entre 2 a 60 kpb. Por exemplo, o transposon composto Tn4001 é constituído pela IS256 da família IS256 e confere resistência aos aminoglicosídeos. Aqui, nós não encontramos a IS256, mas encontramos as IS47, IS128, 174, IS195, 362 e etc., da família IS256 que estavam presentes apenas uma vez nos *contigs.* A dificuldade de montar *contigs* > 5 kpb foi um fator limitante para a recuperação de transposons compostos, e isso pode explicar por que não conseguimos recuperar *contigs* contendo mais de uma IS.

4.8. Busca de integrons

Com o auxílio da ferramenta Integron_finder, encontramos genes cassetes em apenas um *contig* de plasmídeo com tamanho de 2,8 kpb. No entanto, nesse *contig*, não foi encontrado nenhum gene de integrase conhecido. Ao invés disso, encontramos a IS110 associada a uma transposase. Os demais genes correspondiam a proteínas hipotéticas e proteínas sem similaridade com a base de dados do RefSeq. Já em *contigs* de cromossomos que carregavam ISs, encontramos 9 integrons associados com as IS3, ISL3,

IS4, IS481 e IS66, esses *contigs* também tinham tamanho de até 2,8 kpb e carregavam principalmente genes desconhecidos o que impossibilita o esclarecimento de suas funções. Como os *contigs* eram pequenos e carregavam apenas uma IS, não foi possível saber se os integrons estavam inseridos em transposons. Por outro lado, a única evidência de associação entre plasmídeo e integron que encontramos, foi visto no plasmídeo p1522 mostrado na Figura 17.

4.9. Prospecção de ARGs em *contigs* de plasmídeos

O resultado da prospecção de ARGs nos *contigs* cromossômicos e plasmidiais está apresentado na Figura 22. Em *contigs* de plasmídeos conhecidos foram encontrados apenas alguns ARGs incluindo os genes APH(6)-Id e APH(3'')-Ib que conferem resistência aos aminoglicosídeos e o gene sul1 que confere resistência a sulfonamida. Da mesma forma, nos *contigs* de plasmídeos desconhecidos também foram encontrados poucos ARGs, incluindo os genes ANT(3'')-IIb e aadA que também conferem resistência aos aminoglicosídeos, e os genes adeF e tet(L) que conferem resistência a tetraciclina e fluoroquinolona por bombas de efluxo. Nos *contigs* < 1 kpb, também foram encontrados poucos ARGs, ainda que esses *contigs* representem 90% das sequências montadas com MetaSpades. Foram encontrados genes relacionados com a redução da permeabilidade, mutações em alvos do antibiótico como ocorre com o gene de fator de alongamento EF-Tu que conferem resistência a elfamicina, mutação no gene GlpT que confere resistência a fosfomicina, mutações nos genes de rRNA de 16S e 23S que confere resistência a aminoglicosídeos e macrolídeos respectivamente. Foi identificado também resistência a fluoroquinolona por meio da mutação do gene parC que codifica uma proteína subunidade de DNA topoisomerase. Além de outras proteínas de bombas de efluxo mediando resistência a várias classes de antibióticos como fluoroquinolona, cefalosporina, glicilciclina, penam, tetraciclina, rifamicina, fenicol e triclosan. A grande maioria dos ARGs, foram encontrados nos *contigs* de cromossomos com tamanho ≥ 1 kpb incluindo genes da família KPC beta-lactamase. Em resumo, foram encontrados 450 ARGs nos *contigs* dos quais as famílias pertencentes a bombas de efluxo foram as mais abundantes (Figura 22).

4.10. Prospecção de ARGs por mapeamento de *reads*

A prospecção de ARGs também foi realizada nas *reads* dos metagenomas dos diferentes dias da compostagem. Encontramos 836 ARGs dos quais 30% foram preditos pelo modelo variante de proteínas e 70% pelo modelo homologia de proteínas (Figura 23).

O modelo variante está integrado no algoritmo do RGI e consiste na busca de mutações não sinônimas específicas ou outras variantes genéticas como INDELs e *frameshifts* que diferenciam entre alelos do tipo selvagem susceptíveis a antibióticos e alelos resistentes a antibióticos (Alcock *et al*., 2020).

Em relação a variantes de proteínas, foi encontrado resistência contra sulfonamidas pela mutação do gene folP que codifica dihydropteroate synthase, alteração do gene ompF de porina que reduz a permeabilidade de beta-lactamases, mutação nos genes gyrA e gyrB que confere a resistência contra aminocumarina, fluoroquinolona e zoliflodacina, mutação nos genes rpoB e rpoC que codifica subunidades de RNA polimerase conferindo resistência a rifamicina, daptomicina e vancomicina, alteração de proteínas do sistema ABC que medeiam o efluxo de macrolídios e peptídeos, mutação no gene rpsL conferindo resistência a estreptomicina em *Mycobacterium tuberculosis*, alteração da proteína EF-Tu que participa do processo de tradução conferindo resistência a elfamicina, mutação nos genes parC e parE ocasionando a resistência a fluoroquinolonas, mutação do gene CyaA conferindo resistência a fosfomicina, mutação do gene rpsA de resistência a pirazinamida, mutação do gene nfsA de resistência a nitrofurantoína e mutação do gene embC que confere resistência ao etambutol.

Figura 22: Família de ARGs encontrados nos *contigs*.
Chrom. corresponde aos *contigs* cromossômicos; *Contigs* < 1kb corresponde aos *contigs* que não foram usados para a recuperação de plasmídeos; Unk. Plas. corresponde aos *contigs* de plasmídeos desconhecidos; Know Plas. Corresponde aos *contigs* de plasmídeos conhecidos; Qt: corresponde a quantidade de genes por família de AMR.

Na busca pelo modelo de homologia, a maioria dos genes estavam associados com os mecanismos de inativação dos antibióticos e bombas de efluxo, tais como os genes das famílias TEM, OXA e ampC das beta-lactamases, ANT(3''), APH(3''), APH(6) e AAC(6'') das enzimas nucleotidiltransferases, fosfotransferases e também das acetiltransferases respectivamente que atuam inibindo os aminoglicosídeos e proteínas pertencentes aos sistemas ABC, RND, MFS e MATE de bombas de efluxo. Na amostra do dia 01 foram

encontrados a maioria dos genes de resistência, havendo destaque para a família TEM das beta-lactamases e proteínas do sistema RND de bomba de efluxo.

Ao longo de todos os dias houve permanência das famílias de genes de resistência ANT(3''), proteínas de bombas de efluxo, parCE, gyrAB, rpoB, rpsA, ileS e EF-Tu, que estão associados a multirresistência por bomba de efluxo assim com também em casos mais específicos como a resistência a aminoglicosídeos, fluoroquinolonas, zoliflodacina, rifamicina, pirazinamida, mupirocina, elfamicina e aminocumarina.

A abundância de *reads* foi maior nas famílias OXA, APH(3''), APH(6) em proteínas de bomba de efluxo, nas famílias parCE e gyrAB, proteção do ribossomo contra tetraciclina, rpsA e rpsL. Houve um aumento gradual das *reads* ao longo dos dias nas famílias RND e rpoB de resistência a rifamicina. Por outro lado, houve uma diminuição das famílias APH(3''), APH(6), resistência a tetraciclina, aminocumarina e entre outros.

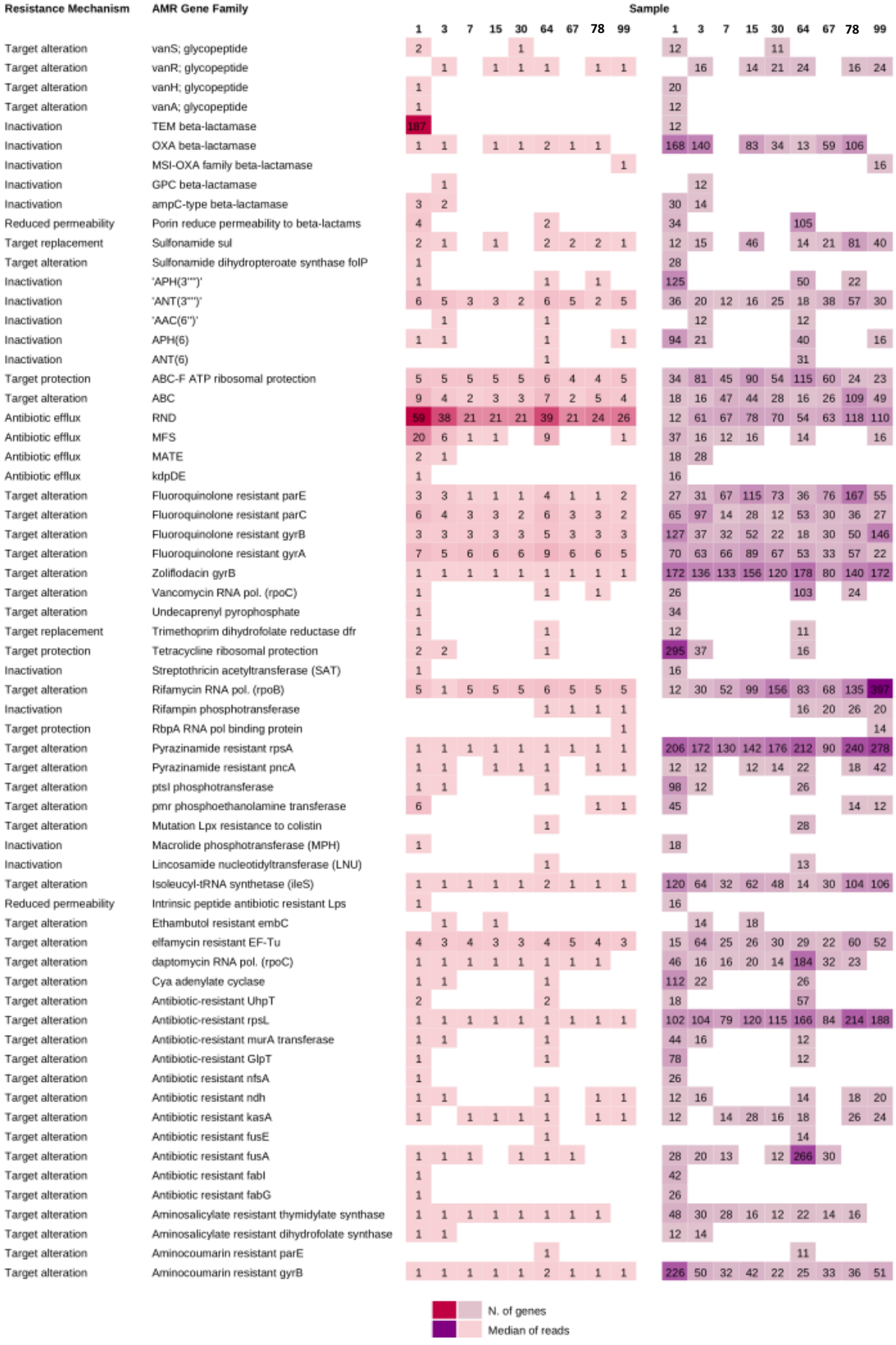


Figura 23: Prospecção de ARGs pelo mapeamento de *reads* do metagenoma *shotgun* de amostras da série temporal do processo de compostagem.

5. DISCUSSÃO

Para montagem dos *contigs* a partir de dados de metagenômica de *short reads* utilizamos o MetaSpades, que é considerado um dos mais eficientes montadores metagenômicos (Lapidus & Korobeynikov, 2021). Dentre as abordagens de co-montagem (mistura de amostras) e montagem individual, escolhemos montar as *reads* das amostras da composteira ZC4 de forma individual, com o propósito de evitar a formação de *contigs* quiméricos originados a partir das diferentes amostras. Porém, ao escolher a montagem individual das amostras, corremos o risco de perder *contigs* menos abundantes nas amostras, já que eles tendem a ter uma baixa cobertura e por isso podem não ser montados (Antipov *et al*., 2019). Por outro lado, se tivéssemos misturado as amostras, poderíamos ter aumentado o número de *reads* das sequências menos abundantes, mas estaríamos também misturando *reads* de cepas microbianas intimamente relacionadas das diferentes amostras. Isso poderia comprometer o processo de montagem porque o grafo de Bruijn poderia ter vários caminhos alternativos. Consequentemente, o montador poderia decidir quebrar o grafo em partes menores, o que poderia resultar na fragmentação dos *contigs* (Delgado & Andersson, 2022). Portanto, na nossa montagem, evitamos a mistura de dados das diferentes amostras, com o propósito de diminuir a montagem de quimeras e a alta fragmentação dos *contigs*. Mesmo assim, em cada amostra tivemos uma grande porcentagem de *contigs* fragmentados (< 1 kpb), que pode ter ocorrido em decorrência da complexidade e grande diversidade genética na comunidade microbiana da compostagem. Quanto mais complexa é a comunidade microbiana, mais recursos computacionais são necessários, além disso, a complexidade dos dados pode dificultar ainda mais o processo de montagem. A alta fragmentação e o aproveitamento de apenas 8,9% dos *contigs* > 1 kpb, foi um fator limitante para a recuperação de plasmídeos completos, entretanto, os resultados que obtivemos foram robustos.

Um estudo comparou a eficiência de várias ferramentas de recuperação de plasmídeos em dados metagenômicos de *short reads* e revelou que de 14% a 50% dos *contigs* pertencem verdadeiramente a plasmídeos (Paganini *et al*., 2021). Isso mostra que ainda existe uma grande porcentagem de *contigs* que são recuperados erroneamente quando a recuperação é realizada com apenas uma ferramenta. Para evitar esse problema, consideramos como *contigs* de plasmídeos apenas aquelas sequências que foram recuperadas por mais de uma ferramenta.

Dentre as várias ferramentas que existem, escolhemos aquelas que foram desenvolvidas principalmente para a recuperação de plasmídeos em dados

metagenômicos, que não sejam limitadas a apenas alguns táxons ou sequências já conhecidas nos banco de dados, e que sejam preferencialmente baseadas em linha de comando, permitindo uma maior autonomia na definição dos parâmetros e que não imponham restrições à quantidade de dados a ser analisado, algo que é limitado nas ferramentas baseadas em interface web.

Com a ferramenta Plasflow (Krawczyk *et al*., 2018) recuperamos a maior quantidade de *contigs* supostamente de plasmídeos. Essa ferramenta leva em consideração a diferença do conteúdo GC entre os plasmídeos e os cromossomos das bactérias hospedeiras pertencentes ao mesmo filo. Em 80% dos plasmídeos, o conteúdo GC é menor do que nos cromossomos dos hospedeiros (Nishida, 2012). É bem provável que exista uma grande variedade de conteúdo GC nos dados da compostagem que analisamos, devido a sua grande complexidade. Isso pode ter contribuído para a classificação de vários *contigs* como sendo de plasmídeos. Um outro ponto a ser discutido, é que o Plasflow além de classificar os plasmídeos, também aponta para qual filo o plasmídeo pertence, tomando como referência o conteúdo GC da hospedeira. Neste sentido, obtivemos 46% dos plasmídeos como pertencentes a hospedeiras do filo Proteobacteria, 13% pertenciam as hospedeiras do filo Firmicutes, 0,6% ao filo da Actinobacteria e 38% correspondiam a plasmídeos que não foi possível identificar a qual filo a hospedeira pertencia. Essa observação concorda com o estudo anterior, que revelou que os membros dos filos Firmicutes, Proteobacteria, Bacteroidetes e Actinobacteria foram as mais abundantes nas mesmas amostras de compostagem que analisamos (Antunes *et al*., 2016). Com isso, supomos que os plasmídeos que encontramos com Plasflow possuem origem principalmente das hospedeiras dos filos Proteobacteria e Firmicutes.

Por outro lado, a ferramenta Plasflow foi a que recuperou a maior quantidade de *contigs* de cromossomos como sendo de plasmídeos. Poderíamos ter aumentado o *threshold* da ferramenta para aumentar sua acurácia, mas levamos em consideração a complexidade dos dados metagenômicos da compostagem e preferimos manter em 0.9 de probabilidade de as sequências serem de plasmídeos. Caso aumentássemos esse valor, poderíamos perder plasmídeos que compartilham características semelhantes com os cromossomos (Antipov *et al*., 2019). No entanto muitos desses *contigs* cromossômicos foram removidos já que não foram recuperados com as demais ferramentas. O mesmo também ocorreu com o MetaPlasmidSpades (Antipov *et al*., 2019) que foi a segunda ferramenta que recuperou o maior número de *contigs* cromossômicos como sendo de plasmídeos, a qual, além de montar os maiores *contigs*, gerou uma grande quantidade de

sequências quiméricas.

O MetaPlasmidSpades (Antipov *et al*., 2019) foi a ferramenta que montou os *contigs* maiores e recuperou a segunda maior quantidade de *contigs* considerados como derivados de plasmídeos. Além disso, essa ferramenta recuperou a maior quantidade de *contigs* com genes de rRNA e genes essenciais. O MetaPlasmidSpades é complementado com as ferramentas ExSPAnder, que busca estender cada caminho no grafo de Bruijn com o propósito de gerar *contigs* mais longos e plasmidVerify que verifica se o *contig* recuperado possui genes característicos de plasmídeos (Antipov *et al*., 2019). Dentre os *contigs* recuperados por MetaPlasmidSpades, existiam aqueles que tinham uma grande porcentagem de genes essenciais, principalmente os *contigs* com tamanho maior que 100 kpb. Na etapa de remoção dos *contigs* com genes rRNA e genes essenciais, esses *contigs* foram removidos, pois suspeitamos que eram *contigs* quiméricos, tinham uma grande porcentagem de genes essenciais e ausência ou casos raros de genes relacionados com a replicação, conjugação e mobilização plasmidial.

Uma explicação provável para os *contigs* maiores e quiméricos montados por MetaPlasmidSpades, é que o exSPAnder ao estender os caminhos no grafo de Bruijn com o propósito de recuperar sequências completas, acaba montando *contigs* maiores do que aqueles montados por MetaSpades. Isso explica, por exemplo, a recuperação de um *contig* com tamanho de 395 kpb com MetaPlasmidSpades, enquanto com MetaSpades o maior *contig* montado foi de 149 kpb.

O quimerismo pode ser explicado pela dificuldade que existe em montar *contigs* a partir de dados altamente complexos como a metagenômica *shotgun* de *short reads*, em que pode ocorrer a montagem de quimeras com maior facilidade do que nos dados de genomas isolados sobre o qual o exSPander foi testado durante o seu desenvolvimento (Prjibelski, *et al*., 2014). Os autores de MetaPlasmidSpades, mostraram que plasmidVerify tem classificado incorretamente vários *contigs* cromossômicos como plasmidiais, e que por si, só, tal ferramenta não é capaz de classificar com precisão os plasmídeos e, portanto, deve ser combinado com outras ferramentas para uma maior precisão (Antipov *et al*., 2019). Isso pode ser uma explicação para a ocorrência da recuperação de vários *contigs* como sendo de plasmídeos, mas que na verdade podem ter sido originados de cromossomos.

Com Platon (Schwengers et al., 2020) recuperamos a terceira maior quantidade de *contigs* como sendo de plasmídeos. A métrica RDS, que verifica a distribuição de genes codificadores de proteínas entre plasmídeos e cromossomos para classificar a origem do *contig*, ainda pode classificar erroneamente os *contigs* tendo em vista que certas proteínas

são codificadas em ambos os *replicons*. Os autores do *pipeline* buscaram melhorar a predição por meio de análise de similaridade contra sequências de referências e pela busca de genes que são bem característicos de plasmídeos (Schwengers *et al*., 2020). Isso permitiu que a ferramenta obtenha um resultado mais acurado. Consequentemente, gerou uma menor quantidade de *contigs* preditos como pertencentes a plasmídeos em comparação às duas ferramentas discutidas anteriormente.

As ferramentas SCAPP (Pellow *et al*., 2021) e Recycler (Rozov *et al*., 2017) recuperaram a menor quantidade de *contigs* como sendo de plasmídeos. O fato é que essas ferramentas buscam recuperar *contigs* cíclicos que representam plasmídeos completos (Pellow et al., 2021). No entanto, a reconstrução de plasmídeos inteiros a partir do sequenciamento de *short reads*, ainda permanece sendo um grande desafio de montagem (Arredondo-Alonso *et al*., 2020). Com essas ferramentas conseguimos recuperar os plasmídeos inteiros, porém todos eles eram pequenos *contigs* circulares.

O processo de anotação dos *contigs* recuperados como sendo de plasmídeos foi uma etapa essencial para a detecção de outros falsos positivos além de *contigs* de cromossomos, tais como as sequências pertencentes a vírus. Por meio da anotação, foram encontradas várias proteínas de vírus, sendo que a ferramenta Platon foi a que mais recuperou os *contigs* relacionados a vírus. Por tanto, a seleção dos *contigs* recuperados com mais de uma ferramenta, foi uma estratégia que mostrou ser eficiente para a seleção dos *contigs* de plasmídeos com maior grau de confiança. A esta estratégia somou-se a anotação e remoção de *contigs* que tinham genes pertencentes a vírus, genes de rRNA e genes essenciais, estes dois últimos considerados como elementos raros em plasmídeos e que normalmente estão mais presentes nos cromossomos (Samuel & Sebastian, 2015). Como consideramos como *contigs* de plasmídeos, somente aqueles que foram recuperados com mais de uma ferramenta, a maior parte dos *contigs* recuperados foram desconsiderados já que foram recuperados por apenas uma ferramenta. A recuperação simultânea de *contigs* com mais de uma ferramenta, aumentou a confiabilidade da nossa prospecção. A estratégia descrita nessa tese é nova, pois não encontramos trabalhos anteriores que levam em consideração a recuperação de plasmídeos com mais de uma ferramenta. E para aumentar ainda mais a confiabilidade na prospecção de *contigs* de plasmídeos levamos em consideração a composição de genes de rRNA e genes essenciais, pois eles são raros em plasmídeos, e geralmente são encontrados nos cromossomos das hospedeiras, tanto é que essa categoria de genes, são usados como marcadores cromossômicos (Anda *et al*., 2015; Samuel & Sebastian, 2015). Os *contigs* com

uma alta densidade de genes essenciais, tais como aqueles envolvidos com a replicação de DNA (dnaA, dnaN, gyrB, gyrA, dnaB, dnaG, polA, dnaE), topoisomerase (topA), RNA polimerase (rpoA, rpoB, rpoC), chaperona (clpX, dnaK, groEL), sistema geral de secreção (Sec, secA, secD, secE, secF, secY, yidC), proteínas ribossômicas, biogênese de tRNA, biossíntese de lipídios, metabolismo do ácido oxocarboxílico (idh1, leuB, leuC, leuD), síntese de mureína, síntese de aminoácidos (triptofano, histidina), síntese de heme (hemB) e entre outros (Tateishi *et al*., 2020), foram desconsiderados como pertencentes a plasmídeos.

A remoção dos *contigs* contendo genes essenciais foi um fator limitante para a recuperação de plasmídeos com tamanho > 100 kpb. Com isso, ficaram de fora também os megaplasmídeos que possuem centenas de milhares de pares de bases, principalmente porque compartilham certas semelhanças com o DNA cromossômico. Além disso, os megaplasmídeos podem evoluir para cromídeos, o que torna ainda mais difícil distinguir um megaplasmídeo de um cromídeo. O baixo número de cópias e a presença de sequências altamente repetidas, também dificultam a recuperação dos megaplasmídeos a partir de dados de *short reads* (Hall *et al*., 2022). O problema das repetições nas sequências que complicam a montagem vem sendo resolvido pelo emprego da tecnologia de sequenciamento de *long reads*. No entanto, essa tecnologia ainda é pouco empregada devido ao seu alto custo, baixa profundidade (número de *reads* gerados por sequenciamento) e maior taxa de erro (Amarasinghe *et al*., 2020).

Como alternativa, para a recuperação de megaplasmídeos a partir de *short reads*, poderíamos considerar o emprego de algoritmos de *machine learning*. Para tal, o algoritmo poderia ser treinado sobre as características que distinguem um megaplasmídeo de um cromídeo. Por exemplo, os cromídeos tendem a se assemelhar mais aos cromossomos, são herdados verticalmente e parecem não ser transmitidos horizontalmente. A perda da transmissão horizontal foi relatada em experimentos de evolução com grandes plasmídeos conjugativos. Outra distinção fundamental é que os cromídeos codificam funções essenciais para a célula. No entanto, assim como os cromídeos, os megaplasmídeos móveis também ocasionalmente carregam genes homólogos que codificam funções essenciais (Hall *et al*., 2022).

O filtro que adotamos para a recuperação dos *contigs* de plasmídeos, nos permitiu selecionar os *contigs* com características mais próximas de plasmídeos. Nesses *contigs* encontramos uma alta quantidade de genes associados com as categorias funcionais X (Mobiloma: profagos, transposons), L (Replicação, recombinação e reparo), V (Mecanismos

de defesa) e K (Transcrição) da base COG. Essas categorias estão relacionadas com a mobilização do DNA, replicação plasmidial, sistema de defesa e regulação transcricional respectivamente. Tais funções são comumente codificadas por genes encontrados em plasmídeos. Isso é uma evidência de que a nossa metodologia é adequada para a recuperação de *contigs* de plasmídeos com uma acurácia maior do que quando é realizada com apenas uma ferramenta, onde existe uma grande porcentagem de sequências cromossômicas sendo recuperadas como sendo de plasmídeos, como ocorreu, por exemplo, com Plasflow e MetaPlasmidSpades.

Para a busca de plasmídeos conhecidos realizamos uma comparação entre os *contigs* recuperados e plasmídeos de referência em bancos de dados. Além disso, realizamos como alternativa, um segundo método de comparação, que consistiu no mapeamento das *reads* contra os plasmídeos em bases de dados. Dentre esses dois métodos, obtivemos um melhor resultado com o mapeamento das *reads*, pois contamos praticamente com todos as *reads* obtidos do sequenciamento, enquanto pela comparação dos *contigs*, contamos com uma quantidade de sequências limitada a apenas aos *contigs* que foram recuperados como pertencentes a plasmídeos.

Dentre os *contigs* de plasmídeos que recuperamos, alguns tiveram uma alta similaridade com plasmídeos já conhecidos no banco de dados do RefSeq. Esses plasmídeos de referência foram encontrados em espécies do gênero *Acinetobacter* e *Escherichia,* ambas pertencentes ao filo Proteobacteria, na espécie *Weissella cibaria* pertencente ao filo Bacillota e espécies *Clostridium perfringens* e *Geobacillus sp*. pertencentes ao filo Firmicutes. Estes resultados juntamente com aqueles obtidos com Plasflow, nos fornecem evidências de que os *contigs* de plasmídeos que recuperamos podem estar presentes em hospedeiros pertencentes a tais grupos taxonômicos. Muitas espécies do gênero *Acinetobacter* podem causar doenças, como *Acinetobacter baumannii* que é responsável por cerca de 80% das infecções respiratórias em humanos. Estudos recentes relataram o surgimento de *A. baumannii* MDR em pacientes hospitalizados imunocomprometidos (Wong *et al*., 2017). *A. baumannii* pode ser encontrado fora do ambiente hospitalar e que MGEs incluindo os plasmídeos carregam ARGs com resistência até mesmo para os antibióticos usados como último recurso como os carbapenêmicos (Jeong *et al*., 2023). Nosso resultado mostra a ocorrências de um plasmídeo termorresistente que pode estar associado a *A. baumannii*. Não encontramos nenhum ARG nesse plasmídeo, mas este pode ser um indicador para a presença de hospedeiras que persistem ao longo da compostagem.

Para a prospecção dos ARGs escolhemos a base de dados CARD (Alcock *et al.*, 2020) e suas ferramentas integradas, por ser um banco de dados curado e estar em constante atualização. No início da análise, tentamos usar também o nanoARG (Arango-Argoty *et al*., 2019) para servir como complemento ao CARD. No entanto, o *pipeline* apenas suporta uma pequena quantidade de *contigs* que podem ser usados como dados de entrada. Isso inviabilizou a análise da grande quantidade de dados que precisávamos analisar. Além disso, os serviços na *web* do nanoARG não estão mais acessíveis.

Realizamos a prospecção de ARGs por dois métodos. O primeiro consistiu na busca dos ARGs nos *contigs* montados com MetaSpades e o segundo consistiu pelo mapeamento das *reads* contra o CARD. Entre esses dois métodos, obtivemos um resistoma mais amplo pelo mapeamento das *reads* que mostrou a presença dos ARGs ao longo da compostagem. Mas somente pelo método de mapeamento, não conseguimos associar a origem dos ARGs, isto é, se pertencem a cromossomos e/ou a plasmídeos. Por outro lado, os *contigs* por serem sequências maiores e terem um maior número de genes, conseguimos associá-los a uma origem com auxílio das ferramentas de predição de plasmídeos. Com isso, a associação dos dois métodos nos permitiu obter um resultado mais completo quanto a prospecção de ARGs nos metagenomas da compostagem.

A maioria dos ARGs estavam presentes nas sequências cromossômicas, nos quais encontramos uma alta quantidade de ARGs associados as bombas de efluxo principalmente a RND e resistência ao trimetoprim por ocorrência de mutação. Os *contigs* < 1kpb representaram 90% das sequências montadas. Nesses *contigs*, encontramos poucos ARGs. Uma explicação para isso é que a maioria dos ARGs conhecidos e depositados no CARD possuem tamanho > 1 kpb (Alcock *et al*., 2020). Nos *contigs* de plasmídeos, encontramos apenas genes de resistência aos aminoglicosídeos, sulfonamidas e tetraciclinas. A presença dos genes ANT(3”), APH(2”), APH(3“), APH(4), APH(6) (aminoglicosídeos), sul1 e sul2 (sulfonamidas) nos *contigs* de plasmídeos, representa uma possibilidade de disseminação desses ARGs na comunidade microbiana da compostagem via conjugação. Durante todo o processo da compostagem verificamos a permanência dos genes parE, parC, gyrA, gyrB (fluoroquinolonas, aminocumarina e zoliflodacina), rpoB (rifamicina) e rpsA (pirazinamida). Estes genes foram detectados pelo modelo variante de proteínas, que corresponde a detecção de mutações nos genes, tais mutações dificultam o reconhecimento e a ação das drogas sobre o alvo molecular.

Os genes vanS, vanH, vanA (glicopeptídeos), TEM, GPC, ampC (β-lactamase), folP (sulfonamida), AAC(3”), ANT(6) (aminoglicosídeos) e rpoC (vancomicina) foram

encontrados principalmente em *contigs* do primeiro dia da compostagem. A permanência de certas hospedeiras de ARGs ao longo do processo da compostagem pode estar ligada à variação de temperatura, umidade, pH, disponibilidade de nutrientes, disponibilidade de oxigênio e entre outros fatores físicos e químicos que moldam o perfil microbiano (Antunes *et al*., 2016; Zhang *et al*., 2020b). A redução de ARGs foi verificada no processo de compostagem de esterco bovino, tendo sido sugerido que a compostagem foi efetiva em reduzir tanto abundância como o tipo de ARG (Zhang *et al.,* 2020b). Nossas análises identificaram que alguns ARGs persistem ao longo do processo enquanto outros se sucedem, como OXA beta-lactamase e a AAC(3") aminoglicosídeo, representando potencial risco de disseminação no meio ambiente pelo uso do composto maduro como fertilizante.

Por fim, através do mapeamento das sequências do metatranscritoma da compostagem nos genes anotados nos *contigs* do conjunto selecionado de 230 plasmídeos, verificamos a expressão de genes envolvidos na mobilização de DNA, tais como genes que codificam transposases, integrases, recombinases e proteínas de conjugação. Também verificamos que genes que codificam para o sistema de defesa tais como resistência a metais pesados, resistência antibióticos e resistência a compostos tóxicos estavam ativos no processo de compostagem.

Testamos a metodologia que estabelecemos para a recuperação de plasmídeos e na prospecção de ARGs de dados metagenômicos de amostras de fezes de macacos bugios (*Alouatta ssp*) criados em cativeiro e bugios de vida livre. Estes dados foram gerados em trabalho do nosso grupo de pesquisa e explorados quanto a composição diversidade microbiana e potencial funcional do microbioma, a partir de *contigs* e MAGs (Franco, 2022). Fizemos a busca de plasmídeos associados aos ARGs nesses dados já que o conteúdo de MGEs e seu potencial funcional ainda não tinham sido analisados. A seguir discutimos nossas observações preliminares dessa análise.

De forma semelhante como ocorreu nos dados dos metagenomas da compostagem ZC4, 82% das *reads* de bugios tiveram boa qualidade de sequenciamento. Por outro lado, nos metagenomas fecais dos bugios, houve menos fragmentação na montagem, pois 14% dos *contigs* tinham o tamanho ≥ 1 Kpb. Foi observado uma maior quantidade de *contigs* de plasmídeos nos bugios de vida livre, mas a quantidade de ARGs foi maior nos *contigs* de plasmídeos de bugios de cativeiro (resultados não mostrados).

Dentre os *contigs* de plasmídeos de bugios de cativeiro, 5 tinham alta similaridade com plasmídeos conhecidos na base de dados do RefSeq, 1 plasmídeo conhecido, foi

encontrado integrado em 3 *contigs* diferentes, o que pode indicar que esse plasmídeo está integrado ao cromossomo da bactéria hospedeira. A maioria dos plasmídeos conhecidos, eram pequenos com tamanho de até 2,1 kpb, somente 1 plasmídeo tinha o tamanho de 60,4 kpb. Por outro lado, nenhum dos *contigs* de plasmídeos de bugios de vida livre, tiveram similaridade aceitável contra a base de dados de plasmídeos. O que pode indicar que os plasmídeos de bugios de vida livre ainda são bastante desconhecidos. De modo geral encontramos mais plasmídeos nos *contigs* de bugios do que em *contigs* da compostagem ZC4.

Em relação aos ARGs, foram encontrados 8 desses genes nos *contigs* de plasmídeos de bugios de cativeiro. Foram eles: aadS, tet(Q), vanT/G/Y/B, e ErmF conferindo resistência a aminoglicosídeos, tetraciclinas, glicopeptídeos, macrolídeos, lincosamidas e estreptograminas. Todos esses ARGs foram detectados pelo modelo de homologia de proteínas e estavam associados aos mecanismos de inativação do antibiótico, proteção da proteína ribossomal e alteração de Erm 23S rRNA metiltransferase. Nos *contigs* de plasmídeos de bugios de vida livre, apenas foram encontrados 2 ARGs, vanT e adeF conferindo respectivamente resistência a glicopeptídeos pelo mecanismo de alteração da molécula alvo e a fluoroquinolonas e tetraciclinas por bomba de efluxo. O gene tet(Q) foi encontrado integrado no plasmídeo pBIF10 (NG_048272.1), um plasmídeo com tamanho de 2,1 kpb, tendo como hospedeira a *Bifidobacterium longum* (Ma et al., 2015). A presença dessa bactéria nos bugios é benéfica para sua saúde, sendo assim, poderia ser utilizada como probiótico em animais de cativeiro (Hernandez-Rodriguez et al., 2019). Na análise realizada por Franco (2022), foi observado a presença de ARGs em MAGs recuperados dos dados de bugios que analisamos. A autora, encontrou o gene Erm associado com *Treponema berlinense* e adeF associado com espécies da família *Elusimicrobiaceae*. Nossa análise ainda preliminar identificou esses genes em plasmídeos o que pode indicar que esses plasmídeos estão associados aos MAGs recuperados previamente (Franco, 2022).

6. CONCLUSÕES E PERSPECTIVAS

Com o surgimento da metagenômica *shotgun* houve o aumento e a disponibilidade pública de um grande volume de dados, principalmente aqueles sequenciados com a plataforma Illumina que geram *short reads* (Grady *et al*., 2016) abrindo oportunidades e desafios para a mineração de dados e possibilitando a compreensão da natureza de um microbioma específico (Liu *et al*., 2022). Neste trabalho nós exploramos dados de

metagenômica de (*short reads*) da compostagem termofílica (Antunes *et al.*, 2016) com o objetivo de investigar o mobiloma (conjunto de MGEs), com ênfase no plasmidoma (conjunto de plasmídeos), e o resistoma (conjunto de ARGs) desse microbioma. Para processar este conjunto, estabelecemos um fluxograma de análises que faz uso de ferramentas computacionais bem-conceituadas, com parâmetros ajustados para nossos objetivos e tipo de dados brutos disponíveis. Consideramos que é muito importante compreender os dados disponíveis, para somente assim poder adotar os melhores ajustes dos parâmetros durante o uso das ferramentas de bioinformática. A metodologia que estabelecemos consistiu no emprego de múltiplas ferramentas com diferentes abordagens para a recuperação de plasmídeos e, desta forma, possibilitou a recuperação mais precisa de *contigs* de plasmídeos a partir de dados de metagenômica *shotgun* de *short reads*, totalizando 230 *contigs* de plasmídeos, em sua ampla maioria novos plasmídeos.

Um dos desafios que ainda persiste, é a reconstrução de genomas a partir de metagenomas complexos quando sequenciados com tecnologias que geram *short reads* (Martin et al., 2020). Com o metagenoma da compostagem também não foi diferente, pois apenas 8,9% dos *contigs* montados tiveram tamanho ≥ 1 kpb, o que pode ter limitado a recuperação de grande parte dos MGEs, já que esses elementos são mais difíceis de serem montados por possuírem sequências repetitivas (Schwengers et al., 2020).

A maioria dos *contigs* de plasmídeo que recuperamos era de *contigs* pequenos constituídos principalmente por genes de transposases, replicação e manutenção plasmidial e sistema de defesa relacionado com a resposta ao estresse ambiental, resistência a metais pesados e a antibióticos das classes dos aminoglicosídeos e sulfonamidas. Grande parte dos *contigs* de plasmídeos que encontramos, tinham ocorrência ao longo do processo de compostagem e na maioria das vezes, essas sequências não foram encontradas no RefSeq, portanto, são potencialmente novos plasmídeos que serão reportados publicamente e serão melhor investigados para possíveis aplicações biotecnológicas como aquelas que requerem plasmídeos contendo genes que conferem resistência a metais pesados e resistência a antibióticos para serem usados como marcadores de resistência em estudos de tecnologia do DNA recombinante. Dos genes encontrados nos *contigs* de plasmídeos, aqueles que codificam transposases e proteínas de replicação e manutenção plasmidial foram os mais mapeados pelas *reads* dos metatranscritomas, indicando que estão expressos na compostagem. Cabe destacar o surgimento de novas bases de dados de plasmídeos por exemplo *The Human Metagenome Assembled Plasmid (MAP) database* (Stockdale *et al.*, 2022) e atualização de bases

existentes como no caso do PLSDB (*Database of Bacterial Plasmids*) (Schmartz *et al.*, 2022), as quais poderão ser futuramente interrogadas para atualização da identificação dos 230 *contigs* de plasmídeos que recuperamos nesse trabalho.

Nossas análises identificaram mais genes de resistência aos antibióticos nos *contigs* de cromossomos dos que nos *contigs* de plasmídeos. Isso pode estar relacionado com a pouca pressão seletiva que ocorre na compostagem diferente do que ocorre em outros ambientes como aqueles oriundos de hospitais onde ocorre uma pressão seletiva maior em decorrência do intenso uso de antibióticos (Wu *et al*., 2022). Obtivemos um maior número de casos de ARGs quando mapeamos as *reads* contra o CARD do que quando fizemos a busca nos *contigs*. A alta fragmentação dos *contigs* pode ter contribuído para a ocorrência de grande número de genes incompletos que não puderam ser preditos. Dessa forma, pelo método de mapeamento recuperamos até mesmo os ARGs que não foram montados completamente nos *contigs* e nos permitiu obter um resistoma mais amplo, com isso, podemos perceber que as bactérias presentes no material da compostagem do Parque Zoológico de São Paulo, são reservatórios ambientais de ARGs. Notamos que alguns dos ARGs ainda permaneceram no final do processo da compostagem que analisamos. Essa observação é de certo modo preocupante, pois o uso do composto maduro como fertilizante pode ser uma fonte de disseminação de ARGs para meio ambiente. Entretanto, é importante ressaltar que o processo de compostagem eliminou boa parte dos ARGs detectados no início do processo.

Com este trabalho estabelecemos uma nova estratégia para a recuperação de sequências plasmidiais a partir de dados de metagenômica shotgun de *short reads.* Temos como perspectivas, contribuir com o depósito em base de dados das novas sequências plasmidiais que encontramos, as quais poderão ser consultadas e melhor investigadas para fins biotecnológicos. Neste trabalho também mostramos o perfil do plasmidoma e do resistoma na compostagem termofílica do Parque Zoológico de São Paulo. A metodologia que estabelecemos poderá ser aplicada em outros conjuntos de dados de metagenômica *shotgun* para a investigação do perfil do plasmidoma e resistoma. Futuramente, outras bases de dados poderão ser associadas a nossa estratégia de recuperação de plasmídeos, permitindo investigar outros plasmídeos. Por fim, baseado em novos testes, nossa estratégia poderá ser ajustada, atualizada e até mesmo automatizada como *pipeline*.

7. REFERÊNCIAS

Ahmed, I., Zhang, Y., Sun, P., Xie, Y. & Zhang, B. 2023. Sensitive response mechanism of ARGs and MGEs to initial designed temperature during swine manure and food waste co-composting. Environmental Research. 216: 114513. Doi:10.1016/j.envres.2022.114513

Alcock, B. P., Raphenya, A. R., Lau, T. T. Y., Tsang, K. K., Bouchard, M. *et al*. 2020. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic Acids Research 48: D517-D525. Doi:10.1093/nar/gkz935

Alic, A. S., Ruzafa, D., Dopazo, J. & Blanquer, I. 2016. Objective review of de novo stand-alone error correction methods for NGS data. WIREs Computational Molecular Science 6:111-146. Doi:10.1002/wcms.1239

Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z. J., Pollard, K. S., Sakharova, E., Parks, D. H., Hugenholtz, P., Segata, N., Kyrpides, N. C. & Finn, R.D. 2021. Unified catalog of 204,938 reference genomes from the human gut microbiome. Nature Biotechnology 39:105–114. Doi:10.1038/s41587-020-0603-3

Altschul, S. F., Gertz, E. M., Agarwala, R., Schaffer, A. A. & Yu, Y. K. 2009. PSI-BLAST pseudocounts and the minimum description length principle. Nucleic Acids Research 37:815-824. Doi:10.1093/nar/gkn981

Amarasinghe, S. L., Su, S., Dong, X. *et al.* 2020 Opportunities and challenges in long-read sequencing data analysis. Genome Biol 21, 30. Doi:10.1186/s13059-020-1935-5

Amgarten, D., Martins, L. F., Lombardi, K. C., Antunes, L. P., de Souza, A. P. S., Nicastro, G. G., Kitajima, E. W., Quaggio, R. B., Upton, C., Setubal, J. C. & da Silva, A. M. 2017. Three novel Pseudomonas phages isolated from composting provide insights into the evolution and diversity of tailed phages. BMC Genomics 18:ARTN 346. Doi:10.1186/s12864-017-3729-z

Anda M, Ohtsubo Y, Okubo T, Sugawara M, Nagata Y, Tsuda M, Minamisawa K, Mitsui H. 2015. Bacterial clade with the ribosomal RNA operon on a small plasmid rather than the chromosome. Proc Natl Acad Sci U S A. 112(46):14343-7. Doi:10.1073/pnas.1514326112.

Andrews, S. 2010. FASTQC. A quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Ansirichaiya, S., *et al*. 2016. “PCR-Based Detection of Composite Transposons and Translocatable Units from Oral Metagenomic DNA.” FEMS Microbiology Letters, Oxford University Press, Sept.

Antimicrobial Resistance Collaborators. 2022. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. Lancet. Feb 12;399(10325):629-655. Doi:10.1016/S0140-6736(21)02724-0

Antipov, D., Raiko, M., Lapidus, A. & Pevzner, P. A. 2019. Plasmid detection and assembly in genomic and metagenomic data sets. Genome Research 29:961-968. Doi:10.1101/gr.241299.118

Antunes, L. P., Martins, L. F., Pereira, R. V., Thomas, A. M., Barbosa, D., Lemos, L. N., Silva, G. M. M., Moura, L. M. S., Epamino, G. W. C., Digiampietri, L. A., Lombardi, K. C., Ramos, P. L., Quaggio, R. B., de Oliveira, J. C. F., Pascon, R. C., da Cruz, J. B., da Silva, A. M. & Setubal, J. C. 2016. Microbial community structure and dynamics in thermophilic composting viewed through metagenomics and metatranscriptomics. Scientific Reports 6:ARTN 38915. Doi:10.1038/srep38915

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H. Z., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N. & Yeh, L. S. L. 2004. UniProt: the Universal Protein knowledgebase. Nucleic Acids Research 32:D115-D119. Doi:10.1093/nar/gkh131

Arango-Argoty, G. A., Dai, D., Pruden, A., Vikesland, P., Heath, L. S. & Zhang, L. 2019. NanoARG: a web service for detecting and contextualizing antimicrobial 29 resistance genes from nanopore-derived metagenomes. Microbiome 7: ARTN 88. Doi:10.1186/s40168-019-0703-9

Arango-Argoty, G., Garner, E., Prudent, A., Heath, L. S., Vikesland, P. & Zhang, L. Q. 2018. DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. Microbiome 6: ARTN 23. Doi:10.1186/s40168-018- 0401-z

Arashida, H., Odake, H., Sugawara, M. *et al.* 2022 .Evolution of rhizobial symbiosis islands through insertion sequence-mediated deletion and duplication. ISME J 16, 112–121. Doi:10.1038/s41396-021-01035-4

Arredondo-Alonso, S., Bootsma, M., Hein, Y., Rogers, M. R. C., Corander, J., Willems, R. J. L. & Schürch, A. C. 2020. Gplas: a comprehensive tool for plasmid analysis using short-read graphs, Bioinformatics, 36:3874–3876, Doi:10.1093/bioinformatics/btaa233

Arredondo-Alonso, S., Rogers, M. R. C., Braat, J. C., Verschuuren, T. D., Top, J., Corander, J., Willems, R. J. L. & Schurch, A. C. 2018. mlplasmids: a user-friendly tool to predict plasmid- and chromosome-derived sequences for single species. Microbial Genomics 4:ARTN 000224. Doi:10.1099/mgen.0.000224

Arredondo-Alonso, S., Willems, R. J., van Schaik, W. & Schurch, A. C. 2017. On the (im) possibility of reconstructing plasmids from whole-genome *short*-read sequencing data. Microbial Genomics 3:ARTN 000128. Doi:10.1099/mgen.0.000128

Babakhani, S. & Oloomi, M. 2018. Transposons: the agents of antibiotic resistance in bacteria. J. Basic Microbiol., 58: 905-917, Doi:10.1002/jobm.201800204

Beauregard, A., Curcio, M. J. & Belfort, M. 2008. The take and give between retrotransposable elements and their hosts. Annual review of genetics 42:587-617. Doi:10.1146/annurev.genet.42.110807.091549

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Ostell, J., Pruitt, K. D. & Sayers, E. W. 2018. GenBank. Nucleic Acids Research 46:D41-D47. Doi:10.1093/nar/gkx1094

Bertelli, C., Tilley, K. E. & Brinkman, F. S. L. 2019. Microbial genomic island. discovery, visualization and analysis. Briefings in Bioinformatics 20:1685-1698. Doi:10.1093/bib/bby042

Bertrand, D., Shaw, J., Kalathiyappan, M. *et al*. 2019. Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. Nat Biotechnol 37, 937–944. Doi:10.1038/s41587-019-0191-2

Bignell, C., Thomas, C. M. 2001. The bacterial ParA-ParB partitioning proteins. J Biotechnol. Sep 13;91(1):1-34. Doi:10.1016/s0168-1656(01)00293-0.

Bishara, A., Moss, E. L., Kolmogorov, M., Parada, A. E., Weng, Z., Sidow, A., Dekas, A. E., Batzoglou, S. & Bhatt, A. S. 2018. High-quality genome sequences of uncultured microbes by assembly of read clouds. Nat Biotechnol. 2018 Oct 15: Doi:10.1038/nbt.4266.\d: 10.1038/nbt.4266

Bolger, A. M., Lohse, M. & Usadel, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114-2120. Doi:10.1093/bioinformatics/btu170

Braga, L. P. P., Pereira, R. V., Martins, L. F., Silva Moura, L. M., Sanchez, F. B., Patané, J. S. L., da Silva, A. M. & Setubal, J. C. 2021. Genome-resolved metagenomics and metatranscriptome analysis reveal key bacterial players and their metabolic interactions in thermophilic composting. Doi:10.1186/s12864-021-07957-9

Brito, I. L., Yilmaz, S., Huang, K., Xu, L., Jupiter, S. D., Jenkins, A. P., Naisilisili, W., Tamminen, M., Smillie, C. S., Wortman, J. R., Birren, B. W., Xavier, R. J., Blainey, P. C., Singh, A. K., Gevers, D. & Alm, E. J. 2016. Mobile genes in the human microbiome are structured from global to individual scales. Nature 535:435. Doi:10.1038/nature18927

Brooks, L., Kaze, M. & Sistrom, M. 2019. A Curated, Comprehensive Database of Plasmid Sequences. Microbiology Resource Announcements 8:ARTN e01325-18. Doi:10.1128/MRA.01325-18

Brown, C. L., Keenum, I. M., Dai, D. *et al*. 2021. Critical evaluation of short, long, and hybrid assembly for contextual analysis of antibiotic resistance genes in complex environmental metagenomes. Sci Rep 11, 3753. Doi:10.1038/s41598-021-83081-8

Buchfink, B., Xie, C. & Huson, D. H. 2015**.** Fast and sensitive protein alignment using DIAMOND. Nature Methods 12:59-60. Doi:10.1038/nmeth.3176

Burrus, V. 2017. Mechanisms of stabilization of integrative and conjugative elements. Current Opinion in Microbiology 38:44-50. Doi:10.1016/j.mib.2017.03.014

Burton, J. N., Liachko, I., Dunham, M. J. & Shendure, J. 2014. Species-level deconvolution of metagenome assemblies with Hi-C-based contact probability maps. G3 (Bethesda). May 22;4(7):1339-46. Doi:10.1534/g3.114.011825.

Bushnell, B. 2014. BBMap: A Fast, Accurate, Splice-Aware Aligner.

Cao, R. K., Bem, W. W., Qiang, Z. M. & Zhang, J. Y. 2020. Removal of antibiotic resistance genes in pig manure composting influenced by inoculation of compound microbial agents Bioresour. Technol., 317.

Carattoli, A., Zankari, E., Garcia-Fernandez, A., Larsen, M. V., Lund, O., Villa, L., Aarestrup, F. M. & Hasman, H. 2014. In Silico Detection and Typing of Plasmids 30 using PlasmidFinder and Plasmid Multilocus Sequence Typing. Antimicrobial Agents and Chemotherapy 58:3895-3903. Doi:10.1128/Aac.02412-14

Castaldi, P., Alberti, G., Merella, R. & Melis, P. 2005. Study of the organic matter evolution during municipal solid waste composting aimed at identifying suitable parameters for the evaluation of compost maturity. Waste Manag. 25(2):209-13. Doi:10.1016/j.wasman.2004.12.011

Chen, L. X., Anantharaman, K., Shaiber, A., Eren, A. M. & Banfield, J. F. 2020. Accurate and complete genomes from metagenomes. Genome Research 30:315- 333. Doi:10.1101/gr.258640.119

Cheng, D., Liu, Y., Shehata, E., Feng, Y., Lin, H., Xue, J. & Li, Z. 2021. In-feed antibiotic use changed the behaviors of oxytetracycline, sulfamerazine, and ciprofloxacin and related antibiotic resistance genes during swine manure composting, Journal of Hazardous Materials, Vol. 402: 123710. Doi:10.1016/j.jhazmat.2020.123710

Claire, H., Christiane, F., Ousmane, T., Didier, D. & Geneviève, B. 2022. Plasmidome analysis of a hospital effluent biofilm: Status of antibiotic resistance. Plasmid. Vol.122, Doi:10.1016/j.plasmid.2022.102638

Clark, D. P., Pazdernik, N. J. & McGehee, M. R. Plasmids. *IN* Molecular Biology, 712–748 (Elsevier, 2019). Doi:10.1016/B978-0-12-813288-3.00023-9.

Craig, L., Forest, K.T. & Maier, B. 2019. Type IV pili: dynamics, biophysics and functional consequences. Nat Rev Microbiol 17, 429–440. Doi:10.1038/s41579-019-0195-4

Cury, J., Jove, T., Touchon, M., Neron, B. & Rocha, E. P. C. 2016. Identification and analysis of integrons and cassete arrays in bacterial genomes. Nucleic Acids Research. Doi:10.1093/nar/gkw319

Cytryn, E. 2013. The soil resistome: The anthropogenic, the native, and the unknown. Soil Biology & Biochemistry 63:18-23. Doi:10.1016/j.soilbio.2013.03.017

Darmon, E. & Leach, D. R. F. Bacterial Genome Instability. 2014. ASM Journals Microbiology and Molecular Biology Reviews Vol. 78, No. 1. Doi:10.1128/MMBR.00035-13

de Carvalho FM, Valiatti TB, Santos FF et al. 2022. Exploring the Bacteriome and Resistome of Humans and Food-Producing Animals in Brazil. Microbiol Spectr. 10(5):e0056522. Doi:10.1128/spectrum.00565-22.

Delgado, L. F. & Andersson, A. F. 2022. Evaluating metagenomic assembly approaches for biome-specific gene catalogues. Microbiome 10, 72. Doi:10.1186/s40168-022-01259-2

Dib, J. R., Wagenknecht, M., Farias, M. E. & Meinhardt, F. 2015. Strategies and approaches in plasmidome studies uncovering plasmid diversity disregarding of linear elements? Frontiers in Microbiology 6:ARTN 433. Doi:10.3389/fmicb.2015.00463

Diene, S. M. & Rolain, J. M. 2014. Carbapenemase genes and genetic platforms in Gram-negative bacilli: Enterobacteriaceae, Pseudomonas and Acinetobacter species. Clinical Microbiology and Infection 20:831-838. Doi:10.1111/1469-0691.12655

Du, J. J., Zhang, Y. Y., Qu, M. X., Yin, Y. T., Fan, K., Hu, B., Zhang, H. Z., Wei, M. B. & Ma, C. 2019. Effects of biochar on the microbial activity and community structure during sewage sludge composting Bioresour. Technol., 272 pp. 171-179. Doi:10.1016/j.biortech.2018.10.020

Dy, R. L, Przybilski, R., Semeijn, K., Salmond, G. P. & Fineran, P. C. 2014. A widespread bacteriophage abortive infection system functions through a Type IV toxin-antitoxin mechanism. *Nucleic Acids Res.* 42, 4590-605

Eisen, J. A. 2007. Environmental *shotgun* sequencing: Its potential and challenges for studying the hidden world of microbes. Plos Biology 5:384-388. Doi:10.1371/journal.pbio.0050082

Endoh, H., Hirayama, T., Aoyama, T., Oka, A. 1990. Characterization of the virA gene of the agropine-type plasmid pRiA4 of Agrobacterium rhizogenes. FEBS Lett. Oct 1;271(1-2):28-32. Doi:10.1016/0014-5793(90)80364-o. PMID: 2226811

Escudero, J. A., Loot, C., Nivina, A. & Mazel, D. 2015. The Integron: Adaptation On Demand. Microbiology Spectrum 3:22. Doi:10.1128/microbiolspec.MDNA3-0019- 2014

Ewels P, Magnusson M, Lundin S, Käller M. 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 32(19):3047-8. Doi:10.1093/bioinformatics/btw354

Ezzariai, A., Hafidi, M., Khadra, A., Aemig, Q., El Felsa, L., Barret, M., Merlina, G., Patureau, D. & Pinelli, E. 2018. Human and veterinary antibiotics during composting of sludge or manure: Global perspectives on persistence, degradation, and resistance genes. Journal of Hazardous Materials 359:465-481. Doi:10.1016/j.jhazmat.2018.07.092

Fan, H., Wu, S., Woodley, J., Zhuang, G., Bai, Z., Xu, S., Wang, X. & Zhuang, X. 2020. Effective removal of antibiotic resistance genes and potential links with archaeal communities during vacuum-type composting and positive-pressure composting. J Environ Sci (China). Mar;89:277-286. Doi:10.1016/j.jes.2019.09.006

Fang, Z., Tan, J., Wu, S., Li, M., Xu, C., Xie, Z. & Zhu, H. 2019. PPR-Meta: a tool for identifying phages and plasmids from metagenomic fragments using deep learning. Gigascience. Jun 1;8(6):giz066. Doi:10.1093/gigascience/giz066

Finn, R. D. 2020. A unified catalog of 204,938 reference genomes from the human gut microbiome. Nature Biotechnology. Doi:10.1038/s41587-020-0603-3

Finn, R. D., Clements, J. & Eddy, S. R. 2011. HMMER web server: interactive sequence similarity searching. Nucleic Acids Research 39:W29-W37. Doi:10.1093/nar/gkr367

Fluit, A. C. & Schmitz, F. J. 1999. Class 1 integrons, gene cassetes, mobility, and epidemiology. European Journal of Clinical Microbiology & Infectious Diseases 18:761-770. Doi:DOI 10.1007/s100960050398

Forsberg, K. J., Patel, S., Gibson, M. K., Lauber, C. L., Knight, R., Fierer, N. & Dantas, G. 2014. Bacterial phylogeny structures soil resistomes across habitats. Nature 509:612-+. Doi:10.1038/nature13377

Forsberg, K. J., Reyes, A., Bin, W., Selleck, E. M., Sommer, M. O. A. & Dantas, G. 2012. The Shared Antibiotic Resistome of Soil Bacteria and Human Pathogens. Science 337:1107-1111. Doi:10.1126/science.1220761

Franco, R. R. A. 2022. Diversidade taxonômica e funcional da microbiota de fezes de macacos bugios (*Alouatta* spp.) de cativeiro e vida livre. Tese de Doutorado. Programa Interunidades de Pós-Graduação em Bioinformática, Universidade de São Paulo, São Paulo, Brasil. 107pp. doi:10.11606/T.95.2022.tde-14022023-130156.

Frost, L. S., Leplae, R., Summers, A. O. & Toussaint, A. 2005. Mobile genetic elements: The agents of open source evolution. Nature Reviews Microbiology 3:722- 732. Doi:10.1038/nrmicro1235

Gajalakshmi, S. & Abbasi, S. A. 2008. Solid Waste Management by Composting: State of the Art. Critical Reviews in Environmental Science and Technology 38:311- 400. Doi:10.1080/10643380701413633

Galata, V., Fehlmann, T., Backes, C. & Keller, A. 2019. PLSDB: a resource of complete bacterial plasmids. Nucleic Acids Research 47:D195-D202. Doi:10.1093/nar/gky1050

Galperin, M. Y., Wolf, Y. I., Makarova, K. S., Vera, A. R, Landsman, D. & Koonin, E. V. 2021. COG database update: focus on microbial diversity, model organisms, and widespread pathogens. Nucleic Acids Res. 49(D1): D274-D281. Doi:10.1093/nar/gkaa1018

Gandhi, N. R., Nunn, P., Dheda, K., Schaaf, H. S., Zignol, M., van Soolingen, D., Jensen, P. & Bayona, J. 2010. Multidrug-resistant and extensively drug-resistant tuberculosis: a threat to global control of tuberculosis. Lancet 375:1830-1843. Doi:10.1016/S0140-6736(10)60410-2

Gao, M., Qiu, T. L., Sun, Y. M. & Wang, X. M. 2018. The abundance and diversity of antibiotic resistance genes in the atmospheric environment of composting plants. Environment International 116:229-238. Doi:10.1016/j.envint.2018.04.028

Garneau-Tsodikova, S. & Labby, K. J. 2016. Mechanisms of resistance to aminoglycoside antibiotics: overview and perspectives. Med chem comm 7:11-27. Doi:10.1039/c5md00344j

Gordon, J. E, Christie, P. J. 2014 The Agrobacterium Ti Plasmids. Microbiol Spectr. 2(6): PLAS-0010-2013. Doi:10.1128/microbiolspec.PLAS-0010-2013

Gou, M., Hu, H. W., Zhang, Y. J., Wang, J. T., Hayden, H., Tang, Y. Q. & He, J. Z. 2018. Aerobic composting reduces antibiotic resistance genes in cattle manure and the resistome dissemination in agricultural soils. Science of the Total Environment 612:1300-1310. Doi:10.1016/j.scitotenv.2017.09.028

Gounot, J. S., Chia, M., Bertrand, D. *et al.*. 2022 Genome-centric analysis of short and long read metagenomes reveals uncharacterized microbiome diversity in Southeast Asians. Nat Commun 13, 6044. Doi:10.1038/s41467-022-33782-z

Grady, M., Heidi, Workentine & Matthew, L. 2016. The Challenge and Potential of Metagenomics in the Clinic. Frontiers in Immunology. Doi:10.3389/fimmu.2016.00029

Grindley, N. D. 2002. The movement of Tn*3*-like elements: transposition and cointegrate resolution, p. 272-302. *In* N. L. Craig, R. Craigie, M. Gellert, and A. M. Lambowitz (ed.), Mobile DNA II. ASM Press, Washington, DC

Grohmann, E., Christie, P. J., Waksman, G. & Backert, S. 2018. Type IV secretion in Gram-negative and Gram-positive bacteria. Molecular Microbiology 107:455-471. Doi:10.1111/mmi.13896

Guima, S. E. S. 2021. Microbial composition of inoculum and mature compost in the São Paulo Zoo Composting process assessed through metagenomics. Dissertação de Mestrado. Programa Interunidades de Pós-Graduação em Bioinformática, Universidade de São Paulo, São Paulo, Brazil. 73pp. Doi:10.11606/D.95.2021.tde-16022021-122522

Guitor, A. K., Raphenya, A. R., Klunk, J., Kuch, M., Alcock, B., Surette, M. G., McArthur, A. G., Poinar, H. N. & Wright, G. D. 2019. Capturing the Resistome: a Targeted Capture Method To Reveal Antibiotic Resistance Determinants in Metagenomes. Antimicrob Agents Chemother. 64(1):e01324-19. Doi:10.1128/AAC.01324-19

Gurevich, A., Saveliev, V., Vyahhi, N., Tesler, G. 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics.;29(8):1072-5. Doi:10.1093/bioinformatics/btt086

Gutierrez, B., Escudero, J. A., Millan, A. S., Hidalgo, L., Carrilero, L., Ovejero, C. M., Santos-Lopez, A., Thomas-Lopez, D. & Gonzalez-Zorn, B. 2012. Fitness Cost and Interference of Arm/Rmt Aminoglycoside Resistance with the RsmF Housekeeping Methyltransferases. Antimicrobial Agents and Chemotherapy 56:2335-2341. Doi:10.1128/Aac.06066-11

Hall, J. P. J., Botelho, J., Cazares, A. & Baltrus, D. A. 2022. What makes a megaplasmid? Philos Trans R Soc Lond B Biol Sci. 377(1842):20200472. Doi:10.1098/rstb.2020.0472

Hall, B. G. Transposable elements as activators of cryptic genes in E. coli. Genetica. 1999;107(1-3):181-7.

Hallstrom, K. N. & McCormick, B. A. 2015. Pathogenicity Islands. In Molecular Medical Microbiology (Second Edition)

Handel, N., Schuurmans, J. M., Feng, Y. F., Brul, S. & ter Kuile, B. H. 2014. Interaction between Mutations and Regulation of Gene Expression during Development of De Novo Antibiotic Resistance. Antimicrobial Agents and Chemotherapy 58:4371-4379. Doi:10.1128/Aac.02892-14 32

Harrison, P. W., Lower, R. P. J., Kim, N. K. D. & Young, J. P. W. 2010. Introducing the bacterial 'chromid': not a chromosome, not a plasmid. Trends Microbiol. 18:141–148

He, Y., Yuan, Q., Mathieu, J., Stadler, L. B., Senehi, N., Sun, R. & Alvarez, P. J. J. 2020. Antibiotic resistance genes from livestock waste: occurrence, dissemination, and treatment NPJ Clean Water, 3 (4):1-11. Doi:10.1038/s41545-020-0051-0

Hegemann, J. D, Birkelbach, J., Walesch, S., Müller, R. 2023 Current developments in antibiotic discovery: Global microbial diversity as a source for evolutionary optimized anti-bacterials. EMBO Rep. 24(1):e56184. Doi:10.15252/embr.202256184

Heritier, C., Poirel, L. & Nordmann, P. 2006. Cephalosporinase over-expression resulting from insertion of ISAba1 in Acinetobacter baumannii. Clinical Microbiology and Infection 12:123-130. Doi:10.1111/j.1469-0691.2005.01320.x

Hernandez-Rodriguez, D., Vasquez-Aguilar, A. A., Serio-Silva, J. C., Rebollar,E. A. & Azaola-Espinosa, A. 2019. Molecular detection of Bifidobacterium spp.in faeces of black howler monkeys (Alouatta pigra). Journal of Medical Primatology 48:99-105. Doi:10.1111/jmp.12395

Hu FP, Guo Y, Zhu DM, *et al.* 2016. Resistance trends among clinical isolates in China reported from CHINET surveillance of bacterial resistance, 2005-2014. Clin Microbiol Infect. Mar;22 Suppl 1:S9-14. Doi:10.1016/j.cmi.2016.01.001

Huan, W., Wilfred, A. & Van der Donk. 2012. Biosynthesis of the Class III Lantipeptide Catenulipeptin. ACS Chem. Biol. 7, 9:1529–1535. Doi:10.1021/cb3002446

Hyatt, D., Chen, G. L., LoCascio, P. F., Land, M. L., Larimer, F. W. & Hauser, L. J. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. Bmc Bioinformatics 11:Artn 119. Doi:10.1186/1471-2105-11-119

Insam, H. & de Bertoldi, M. 2007. Microbiology of the composting process, p. 26-48. In L. F. Diaz, M. de Bertoldi, W. Bidlingmaier & E. Stentiford (ed.), Compost Science and Technology, vol. 1. Elsevier, Amsterdam, The Netherlands.

Jorgensen, T. S., Xu, Z. F., Hansen, M. A., Sorensen, S. J. & Hansen, L. H. 2014. Hundreds of Circular Novel Plasmids and DNA Elements Identified in a Rat Cecum Metamobilome. Plos One 9:ARTN e87924. Doi:10.1371/journal.pone.0087924

Joshua, M., Jonesllana, G., Avigdor, E. & Alan, D. G. (2021). A mobile genetic element increases bacterial host fitness by manipulating development eLife 10:e65924.

Juhas, M., van der Meer, J. R., Gaillard, M., Harding, R. M., Hood, D. W. & Crook, D. W. 2009. Genomic islands: tools of bacterial horizontal gene transfer and evolution. Fems Microbiology Reviews 33:376-393. Doi:10.1111/j.1574-6976.2008.00136.x

Jurado, M., Lopez, M. J., Suarez-Estrella, F., Vargas-Garcia, M. C., LopezGonzalez, J. A. & Moreno, J. 2014. Exploiting composting biodiversity: Study of the persistent and biotechnologically relevant microorganisms from lignocellulose-based composting. Bioresource Technology 162:283-293. Doi:10.1016/j.biortech.2014.03.145

Kolmogorov, M., Yuan, J., Lin, Y. *et al.* 2019. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol 37, 540–546. Doi:10.1038/s41587-019-0072-8

Koonin, E. V. & Wolf, Y. I. 2008. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic Acids Research 36:6688-6719. Doi:10.1093/nar/gkn668

Kothari, A., Wu, Y. W., Chandonia, J. M., Charrier, M., Rajeev, L., Rocha, A. M., Joyner, D. C., Hazen, T. C., Singer, S. W. & Mukhopadhyay, A. 2019. Large circular plasmids from groundwater plasmidomes span multiple incompatibility groups and are enriched in multimetal resistance genes. MBio 10, e02899-e2918.

Krawczyk, P. S., Lipinski, L. & Dziembowski, A. 2018. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. Nucleic Acids Research 46: ARTN e35. Doi:10.1093/nar/gkx1321

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C. & Salzberg, S. L. 2004. Versatile and open software for comparing large genomes. Genome Biology 5:ARTN R12. Doi:10.1186/gb-2004-5-2-r

Langmead, B. & Salzberg, S. L. 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods 9:357-U54. Doi:10.1038/Nmeth.1923

Lanza, V. F., de Toro, M., Garcillan-Barcia, M. P., Mora, A., Blanco, J., Coque, T. M. & de la Cruz, F. 2014. Plasmid Flux in Escherichia coli ST131 Sublineages, Analyzed by Plasmid Constellation Network (PLACNET), a New Method for Plasmid Reconstruction from Whole Genome Sequences. Plos Genetics 10:ARTN e1004766. Doi:10.1371/journal.pgen.1004766

Lemos, L. N., Pereira, R. V., Quaggio, R. B., Martins, L. F., Moura, L. M. S., da Silva, A. R., Antunes, L. P., da Silva, A. M. & Setubal, J. C. 2017. Genome-Centric Analysis of a Thermophilic and Cellulolytic Bacterial Consortium Derived from Composting. Frontiers in Microbiology 8: ARTN 644. Doi:10.3389/fmicb.2017.00644 33

Li, H. & Durbin, R. (2009) Fast and accurate *short* read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60. [PMID: 19451168]

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9. [PMID: 19505943]

Liao, H. P., Lu, X. M., Rensing, C., Friman, V. P., Geisen, S., Chen, Z., Yu, Z., Wei, Z., Zhou, S. G. & Zhu, Y. G. 2018. Hyperthermophilic Composting Accelerates the Removal of Antibiotic Resistance Genes and Mobile Genetic Elements in Sewage Sludge. Environmental Science & Technology 52:266-276. Doi:10.1021/acs.est.7b04483.

Linda van der G. van B., Jaap A. W. & Aldert, L. Z. RFPlasmid: Predicting plasmid sequences from short read assembly data using machine learning bioRxiv 2020.07.31.230631: Doi:10.1101/2020.07.31.23063

Liu, B. & Pop, M. 2009. ARDB--Antibiotic Resistance Genes Database. Nucleic acids research 37:D443-D447. Doi:10.1093/nar/gkn656

Liu, S., Moon, C.D., Zheng, N. et al. Opportunities and challenges of using metagenomic data to bring uncultured microbes into cultivation. Microbiome 10, 76 (2022). https://doi.org/10.1186/s40168-022-01272-5

Lobocka, M. B. Genome of bacteriophage P1. J. Bacteriol. 186, 7032–7068 (2004).

Lopez-Gonzalez, J. A., Lopez, M. J., Vargas-Garcia, M. C., Suarez-Estrella, F., Jurado, M. & Moreno, J. 2013. Tracking organic matter and microbiota dynamics during the stages of lignocellulosic waste composting. Bioresource Technology 146:574-584. Doi:DOI 10.1016/j.biortech.2013.07

Lopez-Gonzalez, J. A., Suarez-Estrella, F., Vargas-Garcia, M. C., Lopez, M. J., Jurado, M. M. & Moreno, J. 2015. Dynamics of bacterial microbiota during lignocellulosic waste composting: Studies upon its structure, functionality and biodiversity. Bioresource Technology 175:406-416. Doi:10.1016/j.biortech.2014.10

Ma, Y., Xie, T. T., Hu, Q., Qiu, Z., Song, F. Sequencing analysis and characterization of the plasmid pBIF10 isolated from Bifidobacterium longum. Can J Microbiol. 2015 Feb;61(2):124-30. Doi:10.1139/cjm-2014-0581

Maguire, F., Jia, B. F., Gray, K. L., Lau, W. Y. V., Beiko, R. G. & Brinkman, F. S. L. 2020. Metagenome-assembled genome binning methods with *short reads* disproportionately fail for plasmids and genomic Islands. Microbial Genomics 6:ARTN 000436. Doi:10.1099/mgen.0.000436

Maneewannakul, S., Maneewannakul, K. & Ippen-Ihler, K. Characterization, localization, and sequence of F transfer region products: the pilus assembly gene product TraW and a new product, TrbI. Journal of Bacteriology. 1992 Sep;174(17):5567-5574. DOI:10.1128/jb.174.17.5567-5574.1992. PMID: 1355084; PMCID: PMC206500.

Martin, A., Matthew, D. C., & Richard, M. L. New approaches for metagenome assembly with *short reads*, Briefings in Bioinformatics, Volume 21, Issue 2, March 2020, Pages 584–594, https://doi.org/10.1093/bib/bbz020

Martins, L. F., Antunes, L. P., Pascon, R. C., de Oliveira, J. C., Digiampietri, L. A., Barbosa, D., Peixoto, B. M., Vallim, M. A., Viana-Niero, C., Ostroski, E. H., Telles, G. P., Dias, Z., da Cruz, J. B., Juliano, L., Verjovski-Almeida, S., da Silva, A. M. & Setubal, J. C. 2013. Metagenomic analysis of a tropical composting operation at the Sao Paulo Zoo park reveals diversity of biomass degradation functions and organisms. PLoS One 8:e61928. Doi:10.1371/journal.pone.0061928

McArthur, A. G., Waglechner, N., Nizam, F., Yan, A., Azad, M. A. *et al.*. 2013. The Comprehensive Antibiotic Resistance Database. Antimicrobial Agents and Chemotherapy 57:3348-3357. Doi:10.1128/aac.00419-13

McKinney, C. W., Dungan, R. S., Moore, A. & Leytem, A. B. 2018. Occurrence and abundance of antibiotic resistance genes in agricultural soil receiving dairy manure. Fems Microbiology Ecology 94. Doi:10.1093/femsec/fiy010

Mediavilla, J. R., Patrawalla, A., Chen, L., Chavda, K. D., Mathema, B., Vinnard, C., Dever, L. L. & Kreiswirth, B. N. 2016. Colistin- and Carbapenem-Resistant Escherichia

coli Harboring mcr-1 and bla(NDM-5), Causing a Complicated Urinary Tract Infection in a Patient from the United States. Mbio 7:ARTN e01191. Doi:10.1128/mBio.01191-16

Mikheenko, A., Saveliev, V. & Gurevich, A. MetaQUAST: evaluation of metagenome assemblies, Bioinformatics, Volume 32, Issue 7, April 2016, Pages 1088–1090, Doi:10.1093/bioinformatics/btv697

Mirzaei, B., Bazgir, Z. N., Goli, H. R., Iranpour, F., Mohammadi, F. & Babaei, R. 2020. Prevalence of multi-drug resistant (MDR) and extensively drug-resistant (XDR) phenotypes of Pseudomonas aeruginosa and Acinetobacter baumannii isolated in clinical samples from Northeast of Iran. BMC Research Notes 13:380. Doi:10.1186/s13104-020-05224-w

Monzingo, A. F., Ozburn, A., Xia, S., Meyer, R. J. & Robertus, J. D. The structure of the minimal relaxase domain of MobA at 2.1 A resolution. J Mol Biol. 2007 Feb 9;366(1):165-78. Doi:10.1016/j.jmb.2006.11.031. Epub 2006 Nov 11. PMID: 17157875; PMCID: PMC1894915.

MLA style: Barbara McClintock – Facts. NobelPrize.org. Nobel Prize Outreach AB 2023. Wed. 1 Nov 2023. https://www.nobelprize.org/prizes/medicine/1983/mcclintock/facts.

Müller, R. & Chauve, C. HyAsP, a greedy tool for plasmids identification. Bioinformatics. 2019 Nov 1;35(21):4436-4439. Doi:10.1093/bioinformatics/btz413.

Nayfach, S., Roux, S., Seshadri, R. et al. A genomic catalog of Earth's microbiomes. Nature Biotechnology 39:499–509 (2021). Doi:10.1038/s41587-020-0718-6

Nayfach, S., Roux, S., Seshadri, R., Udwary, D., Varghese, N. et al. 2020. A genomic catalog of Earth's microbiomes. Nature Biotechnology. Doi:10.1038/s41587-020-0718-6

Nazir, A., Zhao, Y., Li, M., Manzoor, R., Tahir, R. A., Zhang, X., Qing, H. & Tong Y. Structural Genomics of repA, repB1-Carrying IncFIB Family pA1705-qnrS, P911021-tetA, and P1642-tetA, Multidrug-Resistant Plasmids from Klebsiella pneumoniae. Infect Drug Resist. 2020;13:1889-1903 https://doi.org/10.2147/IDR.S228704

Nicolas, E., Lambin, M., Dandoy, D., Galloy, C., Nguyen, N., Oger, C. A. & Hallet, B. The Tn3-family of Replicative Transposons. ASM Journals Microbiology Spectrum.Vol. 3, No. 4. 23 July 2015. Doi:10.1128/microbiolspec.MDNA3-0060-2014

Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. 2017. metaSPAdes: a new versatile metagenomic assembler. Genome Research 27:824-834. Doi:10.1101/gr.213959.116

Ounit, R., Wanamaker, S., Close, T. J. & Lonardi, S. 2015. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. Bmc Genomics 16:ARTN 236. Doi:10.1186/s12864-015-1419-2

Paganini, J. A., Plantinga, N. L., Arredondo-Alonso, S., Willems, R.J. L. & Schürch, A. C. Recovering Escherichia coli Plasmids in the Absence of Long-Read Sequencing Data. Microorganisms 2021, 9, 1613. Doi:10.3390/microorganisms9081613

Page, R. & Peti, W. Toxin–antitoxin systems in bacterial growth arrest and persistence. Nat. Chem. Biol. 12, 208–214 (2016).

Partanen, P., Hultman, J., Paulin, L., Auvinen, P. & Romantschuk, M. 2010. Bacterial diversity at different stages of the composting process. BMC microbiology 10. Doi:10.1186/1471-2180-10-94

Partridge, S. R., Kwong, S. M., Firth, N. & Jensen, S. O. 2018. Mobile Genetic Elements Associated with Antimicrobial Resistance. Clinical Microbiology Reviews 31:UNSP e00088-17. Doi:10.1128/CMR.00088-17

Partridge, S. R., Tsafnat, G., Coiera, E. & Iredell, J. R. 2009. Gene cassetes and cassete arrays in mobile resistance integrons. Fems Microbiology Reviews 33:757- 784. Doi:10.1111/j.1574-6976.2009.00175.x

Pearson, W. R. (2013). An introduction to sequence similarity ("homology") searching. Current protocols in bioinformatics, Chapter 3, Unit3.1. https://doi.org/10.1002/0471250953.bi0301s42

Pellow, D., Mizrahi, I. & Shamir, R. PlasClass improves plasmid sequence classification. PLoS Comput Biol. 2020 Apr 3;16(4):e1007781. Doi:10.1371/journal.pcbi.1007781.

Pellow, D., Zorea, A., Probst, M., Furman, O., Segal, A., Mizrahi, I. & Shamir, R. SCAPP: an algorithm for improved plasmid assembly in metagenomes. Microbiome. 2021 Jun 25;9(1):144. Doi:10.1186/s40168-021-01068-z. PMID: 34172093; PMCID: PMC8228940.

Penders, J., Stobberingh, E. E., Savelkoul, P. H. M. & Wolffs, F. G. 2013. The human microbiome as a reservoir of antimicrobial resistance. Frontiers in Microbiology 4:ARTN 87. Doi:10.3389/fmicb.2013.00087

Perez, M. F., Saona, L. A., Farías, M. E. *et al.*. Assessment of the plasmidome of an extremophilic microbial community from the Diamante Lake, Argentina. Sci Rep 11, 21459 (2021). https://doi.org/10.1038/s41598-021-00753-1

Perez-Cobas, A. E., Gomez-Valero, L. & Buchrieser, C. 2020. Metagenomic approaches in microbial ecology: an update on whole-genome and marker gene 35 sequencing analyses. Microbial Genomics 6:ARTN 000409. Doi:10.1099/mgen.0.000409

Pfeifer E, Bonnin RA, Rocha EPC. 2022. Phage-Plasmids Spread Antibiotic Resistance Genes through Infection and Lysogenic Conversion. mBio. 13(5):e0185122. Doi:10.1128/mbio.01851-22

Qi Q, Rajabal V, Ghaly TM, Tetu SG, Gillings MR. 2023. Identification of integrons and gene cassette-associated recombination sites in bacteriophage genomes. Front Microbiol. 14:1091391. Doi:10.3389/fmicb.2023.1091391

Queenan, A. M. & Bush, K. 2007. Carbapenemases: the versatile beta-lactamases. Clinical Microbiology Reviews 20:440-458. Doi:10.1128/cmr.00001-07

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. 2017. *Shotgun* metagenomics, from sampling to analysis. Nature Biotechnology 35:833- 844. Doi:10.1038/nbt.3935

Radstrom, P., Skold, O., Swedberg, G., Flensburg, J., Roy, P. H. & Sundstrom, L. 1994. Transposon Tn5090 of plasmid R751, which carries an integron, is related to Tn7, Mu, and the retroelements. J Bacteriol 176:3257–3268.

Ramos PL, Kondo MY, Santos SMB, de Vasconcellos SP, Rocha RCS, da Cruz JB, Eugenio PFM, Cabral H, Juliano MA, Juliano L, Setubal JC, da Silva AM, Cappelini LTD. A Tropical Composting Operation Unit at São Paulo Zoo as a Source of Bacterial Proteolytic Enzymes. Appl Biochem Biotechnol. 2019 187(1):282-297. Doi:10.1007/s12010-018-2810-7

Ramsay, J. P. & Firth, N. 2017. Diverse mobilization strategies facilitate transfer of non conjugative mobile genetic elements. Current Opinion in Microbiology 38:1-9. Doi:10.1016/j.mib.2017.03.003

Ravindran S. 2012. Barbara McClintock and the discovery of jumping genes. Proc Natl Acad Sci U S A. Dec 11;109(50):20198-9. Doi:10.1073/pnas.1219372109

Revilla, C., Garcillan-Barcia, M. P., Fernandez-Lopez, R., Thomson, N. R., Sanders, M., Cheung, M., Thomas, C. A. & de la Cruz, F. 2008. Different pathways to acquiring resistance genes illustrated by the recent evolution of IncW plasmids. Antimicrobial Agents and Chemotherapy 52:1472-1480. Doi:10.1128/Aac.00982-07

Reygaert, W. C. 2018. An overview of the antimicrobial resistance mechanisms of bacteria. AIMS Microbiol 4:482-501. Doi:10.3934/microbiol.2018.3.482

Rodríguez-Beltrán, J., Sørum, V., Toll-Riera, M. & Millán, A. S. 2020 Genetic dominance governs the evolution and spread of mobile genetic elements in bacteria.. 117 (27) 15755-15762. Doi:10.1073/pnas.2001240117

Rozov, R., Kav, A. B., Bogumil, D., Shterzer, N., Halperin, E., Mizrahi, I. & Shamir, R. 2017. Recycler: an algorithm for detecting plasmids from de novo assembly graphs. Bioinformatics 33:475-482. Doi:10.1093/bioinformatics/btw651

Ryan, R. W., Louise, M. J., Claire, L. G. & Kathryn, E. H. 2017. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. Doi:10.1371/journal.pcbi.1005595

Rynk, R. *et al.* 1992. On-Farm Composting Handbook, Northeast Regional Agricultural Engineering Service – Cooperative Extension, Ed., NRAES-54, 185 pag., (Northeast Regional Agricultural Engineering Service, 1992).

Saak, C. C., Dinh, C. B. & Dutton, R. J. 2020. Experimental approaches to tracking mobile genetic elements in microbial communities. Fems Microbiology Reviews 44:606-630. Doi:10.1093/femsre/fuaa025

Samuel, J. Tazzyman & Sebastian Bonhoeffer. 2015. Why There Are No Essential Genes on Plasmids, Molecular Biology and Evolution. (32)12:3079–3088. Doi:10.1093/molbev/msu293

Schmartz GP, Hartung A, Hirsch P, Kern F, Fehlmann T, Müller R, Keller A. 2022. PLSDB: advancing a comprehensive database of bacterial plasmids. Nucleic Acids Res. 50(D1):D273-D278. Doi:10.1093/nar/gkab1111

Scholz, M. RPKM calculation. https://www.metagenomics.wiki/pdf/qc/RPKM. 2022.

Schwengers, O., Barth, P., Falgenhauer, F. L., Hain, T., Chakraborty, T. & Goesmann, A. 2020. Platon: identification and characterization of bacterial plasmid *contigscontigs* in *short*-read draft assemblies exploiting protein sequence-based replicon distribution scores. Microb Genom. 6(10): mgen000398. Doi:10.1099/mgen.0.000398

Seemann, T. 2013. Barrnap 0.7: rapid ribosomal RNA prediction. https://vicbioinformatics.com/software.barrnap.shtml

Siguier, P., Gourbeyre, E., Varani, A., Bao, T. H. & Chandler, M. 2015. Everyman's Guide to Bacterial Insertion Sequences. Microbiology Spectrum 3:ARTN MDNA3- 0030-2014. Doi:10.1128/microbiolspec.MDNA3-0030-2014

Sitaraman, R. 2018. Prokaryotic horizontal gene transfer within the human holobiont: ecological-evolutionary inferences, implications and possibilities. Microbiome 6. Doi:10.1186/s40168-018-0551-z

Skold, O. 2001. Resistance to trimethoprim and sulfonamides. Veterinary Research 32:261-273. Doi:10.1051/vetres:2001123

Smillie, C., Garcillán-Barcia, M. P., Francia, M. V., Rocha, E. P. & de la Cruz F. 2010. Mobility of plasmids. Microbiol Mol Biol Rev. 74(3):434-52. Doi:10.1128/MMBR.00020-10

Snesrud, E., He, S., Chandler, M., Dekker, J. P., Hickman, A. B., McGann, P. & Dyda, F. 2016. A Model for Transposition of the Colistin Resistance Gene mcr-1 by ISApl1. Antimicrobial Agents and Chemotherapy 60:6973-6976. Doi:10.1128/Aac.01457-16 36

Soucy, S. M., Huang, J. L. & Gogarten, J. P. 2015. Horizontal gene transfer: building the web of life. Nature Reviews Genetics 16:472-482. Doi:10.1038/nrg3962

Stockdale, S.R., Harrington, R.S., Shkoporov, A.N. *et al*. 2022. Metagenomic assembled plasmids of the human microbiome vary across disease cohorts. Sci Rep 12, 9212. Doi:10.1038/s41598-022-13313-y

Su, J. Q., Wei, B., Ou-Yang, W. Y., Huang, F. Y., Zhao, Y., Xu, H. J. & Zhu, Y. G. 2015. Antibiotic Resistome and Its Association with Bacterial Communities during Sewage Sludge Composting. Environmental Science & Technology 49:7356-7363. Doi:10.1021/acs.est.5b01012

Suzuki, Y., Nishijima, S., Furuta, Y. *et al.* 2019. Long-read metagenomic exploration of extrachromosomal mobile genetic elements in the human gut. Microbiome 7, 119. Doi:10.1186/s40168-019-0737-z

Tansirichaiya S, Mullany P, Roberts AP. 2016. PCR-based detection of composite transposons and translocatable units from oral metagenomic DNA. FEMS Microbiol Lett. 363(18):fnw195. Doi:10.1093/femsle/fnw195

Tateishi, Y., Minato, Y., Baughn, A. D. *et al.* 2020. Genome-wide identification of essential genes in Mycobacterium intracellulare by transposon sequencing — Implication for metabolic remodeling. Sci Rep 10, 5449. Doi:10.1038/s41598-020-62287-2

Tatusova, T., Ciufo, S., Federhen, S., Fedorov, B., McVeigh, R., O'Neill, K., Tolstoy, I. & Zaslavsky, L. 2015. Update on RefSeq microbial genomes resources. Nucleic Acids Research 43:D599-D605. Doi:10.1093/nar/gku1062

Tsushima, A., Gan, P., Kumakura, N., Narusaka M., Takano, Y., Narusaka, Y. & Shirasu K. 2019. Genomic Plasticity Mediated by Transposable Elements in the Plant Pathogenic Fungus Colletotrichum higginsianum, Genome Biology and Evolution, Volume 11, Issue 5, Pages 1487–1500. Doi:10.1093/gbe/evz087

Van Duin, D. & Paterson, D. L. 2016. Multidrug-Resistant Bacteria in the Community Trends and Lessons Learned. Infectious Disease Clinics of North America 30:377. Doi:10.1016/j.idc.2016.02.004

Van Goethem, M. W., Pierneef, R., Bezuidt, O. K. I., De Peer, Y. V., Cowan, D. A. & Makhalanyane, T. P. 2018. A reservoir of 'historical' antibiotic resistance genes in remote pristine Antarctic soils. Microbiome 6. Doi:10.1186/s40168-018-0424-5

Van Rossum, G. & Drake, F. L. 2009. Python 3 Reference Manual. Scotts Valley, CA: CreateSpace

Vetting, M. W., Hegde, S. S., Wang, M. H., Jacoby, G. A., Hooper, D. C. & Blanchard, J. S. 2011. Structure of QnrB1, a Plasmid-mediated Fluoroquinolone Resistance Factor. Journal of Biological Chemistry 286:25265-25273. Doi:10.1074/jbc.M111.226936

Vielva, L., de Toro, M., Lanza, V. F. & de la Cruz, F. 2017. PLACNETw: a webbased tool for plasmid reconstruction from bacterial genomes. Bioinformatics 33:3796-3798. Doi:10.1093/bioinformatics/btx462

Von Wintersdorff, C. J. H., Penders, J., van Niekerk, J. M., Mills, N. D., Majumder, S., Van Alphen, L. B., Savelkoul, P. H. M. & Wolffs, P. F. G. 2016. Dissemination of Antimicrobial Resistance in Microbial Ecosystems through Horizontal Gene Transfer. Frontiers in Microbiology 7:ARTN 173. Doi:10.3389/fmicb.2016.00173

Walker, A. 2012. Welcome to the plasmidome. Nature Reviews Microbiology 10:379- 379. Doi:10.1038/nrmicro2804

Wallden, K., Rivera-Calzada, A. & Waksman, G. 2010. Type IV secretion systems: versatility and diversity in function. Cell Microbiol. 12(9):1203-12. Doi:10.1111/j.1462-5822.2010.01499.x.

Wang, C., Dong, D., Strong, P. J., Zhu, W. J., Ma, Z., Qin, Y. & Wu, W. X. 2017. Microbial phylogeny determines transcriptional response of resistome to dynamic composting processes. Microbiome 5. Doi:10.1186/s40168-017-0324-0

Weizhong, L. & Adam, G. 2006.Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. Volume 22. Issue 13, Pages 1658–1659. Doi:10.1093/bioinformatics/btl158

WHO, 2020, posting date. Antimicrobial resistance/World Health Organization. https://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance. [Online.] 111.

Wong, D., Nielsen, T. B, Bonomo, R. A. *et al.* 2017**.** Clinical and pathophysiological overview of Acinetobacter infections: A century of challenges. Clin Microbiol Rev 30(1):409–447. Doi:10.1128/CMR.00058-16

Wozniak, R. & Waldor, M. 2010. Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. Nat Rev Microbiol 8, 552–563. Doi:10.1038/nrmicro2382

Wu, D., Jin, L., Xie, J. *et al.* 2022. Inhalable antibiotic resistomes emitted from hospitals: metagenomic insights into bacterial hosts, clinical relevance, and environmental risks. Microbiome **10**, 19. Doi:10.1186/s40168-021-01197-5

Xiao, T. T. & Zhou, W. H. 2020. The third generation sequencing: the advanced approach to genetic diseases. Translational Pediatrics 9:163-173. Doi:10.21037/tp.2020.03.06

Xie, W. Y., Yang, X. P., Li, Q., Wu, L. H., Shen, Q. R. & Zhao, F. J. 2016. Changes in antibiotic concentrations and antibiotic resistome during commercial composting of animal manures. Environmental Pollution 219:182-190. Doi:10.1016/j.envpol.2016.10.044

Xie, Z. & Tang, H. 2017. ISEScan: automated identification of insertion sequence elements in prokaryotic genomes. Bioinformatics. 33(21):3340-3347. Doi:10.1093/bioinformatics/btx433

Yang, Y., Xu, G., Liang, J. *et al*. 2017. DNA Backbone Sulfur-Modification Expands Microbial Growth Range under Multiple Stresses by its anti-oxidation function. Sci Rep 7, 3516. Doi:10.1038/s41598-017-02445-1

Youngblut, N. D., De la Cuesta-Zuluaga, J., Reischer, G. H., Dauser, S., Schuster, N., Walzer, C., Stalder, G., Farnleitner, A. H. & Ley, R. E. 2020. Large Scale Metagenome Assembly Reveals Novel Animal-Associated Microbial Genomes, Biosynthetic Gene Clusters, and Other Genetic Diversity. Msystems 5:ARTN e01045- 20. Doi:10.1128/mSystems.01045-20

Yue, Y., Huang, H., Qi, Z., Dou, H.-M., Liu, X.-Y., Han, T.-F., Chen, Y., Song, X.-J., Zhang, Y.-H. & Tu, J. 2020. Evaluating metagenomics tools for genome binning with real metagenomic datasets and CAMI datasets. BMC Bioinformatics 21:334. Doi:10.1186/s12859-020-03667-3

Zhang, H. W., Jain, C. & Aluru, S. 2020a. A comprehensive evaluation of long read error correction methods. Bmc Genomics 21:15. Doi:10.1186/s12864-020-07227-0

Zhang, L., Gu, J., Wang, X. J., Sun, W., Yin, Y. A., Sun, Y. X., Guo, A. Y. & Tuo, X. X. 2017. Behavior of antibiotic resistance genes during co-composting of swine manure with

Chinese medicinal herbal residues. Bioresource Technology 244:252- 260. Doi:10.1016/j.biortech.2017.07.035

Zhang, M., He, L.-Y., Liu, Y.-S., Zhao, J.-L., Zhang, J.-N., Chen, J., Zhang, Q.-Q. & Ying, G.-G. 2020b. Variation of antibiotic resistome during commercial livestock manure composting. Environment International 136:105458. Doi:10.1016/j.envint.2020.105458

Zhang, T., Yang, Y. & Pruden, A. 2015. Effect of temperature on removal of antibiotic resistance genes by anaerobic digestion of activated sludge revealed by metagenomic approach. Applied Microbiology and Biotechnology 99:7771-7779. Doi:10.1007/s00253-015-6688-9

Zhou, F. F. & Xu, Y. 2010. cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. Bioinformatics 26:2051-2052. Doi:10.1093/bioinformatics/btq299

Zhou, X., Qiao, M., Su, J. Q., Wang, Y., Cao, Z. H., Cheng, W. D. & Zhu, Y. G. 2019. Turning pig manure into biochar can effectively mitigate antibiotic resistance genes as organic fertilizer. Science of the Total Environment 649:902-908. Doi:10.1016/j.scitotenv.2018.08.368

Zhu, W. H., Lomsadze, A. & Borodovsky, M. 2010. Ab initio gene identification in metagenomic sequences. Nucleic Acids Research 38:ARTN e132. Doi:10.1093/nar/gkq275

Zou, D., Ma, L. N., Yu, J. & Zhang, Z. 2015. Biological Databases for Human Research. Genomics Proteomics & Bioinformatics 13:55-63. Doi:10.1016/j.gpb.2015.01.006

8. ANEXOS

Anexo 1. Lista dos 230 *contigs* plasmidiais filtrados manualmente.

circular	tam.*contig* (pb)	*contig*
Yes	1033	99pwNODE_34397_length_1033_cov_16.842163
Yes	1063	78plNODE_31797_length_1063_cov_1.939103
Yes	1147	64pwNODE_27497_length_1147_cov_2.441176
Yes	1231	64plNODE_23343_length_1231_cov_1.466486
Yes	1292	64mNODE_1055_length_1292_cov_2.065236_component_638
Yes	1304	67mNODE_296_length_1304_cov_2.468139_component_173
Yes	1304	78mNODE_855_length_1304_cov_1.994902_component_547
yes	1388	67plNODE_5135_length_1388_cov_1.693101
yes	1404	7mNODE_440_length_1404_cov_3.866875_component_123
yes	1407	15mNODE_457_length_1407_cov_5.360938_component_209
yes	1465	64pwNODE_15514_length_1465_cov_2.652466
yes	1515	78plNODE_15994_length_1515_cov_1.423631
yes	1544	1plNODE_9612_length_1544_cov_10.550459
yes	1546	3plNODE_11299_length_1546_cov_3.441156
yes	1566	64pwNODE_13383_length_1566_cov_1.836692
yes	1579	64plNODE_13151_length_1579_cov_2.687328
yes	1595	3mNODE_697_length_1595_cov_5.094687_component_440
yes	1616	67pwNODE_3516_length_1616_cov_3.490262
yes	1630	30plNODE_6286_length_1630_cov_1.216234
yes	1648	3pw03NODE_10054_length_1648_cov_4.058514
yes	1648	15pwNODE_6936_length_1648_cov_6.863905
yes	1649	7plNODE_7991_length_1649_cov_8.212221
yes	1702	99mNODE_725_length_1702_cov_11.055873_component_568
yes	1709	67pwNODE_3071_length_1709_cov_5.223767
yes	1723	15pwNODE_6272_length_1723_cov_3.781328
yes	1752	1pwNODE_7253_length_1752_cov_10.755692
yes	1752	15pwNODE_6049_length_1752_cov_44.489231
yes	1752	30pwNODE_5226_length_1752_cov_11.286769
yes	1752	3mNODE_679_length_1752_cov_16.504615_component_438
yes	1752	7mNODE_396_length_1752_cov_46.145846_component_118
yes	1752	64mNODE_991_length_1752_cov_45.847385_component_631
yes	1752	67mNODE_271_length_1752_cov_33.940308_component_168
yes	1800	99plNODE_9355_length_1800_cov_11.285714
yes	1804	99plNODE_9306_length_1804_cov_2.520572
yes	1831	64pwNODE_9456_length_1831_cov_5.951878
yes	1833	67mNODE_268_length_1833_cov_5.685229_component_167
yes	1846	64plNODE_9289_length_1846_cov_4.418266
yes	1852	78mNODE_785_length_1852_cov_4.105507_component_544
yes	1854	67mNODE_266_length_1854_cov_32.613781_component_165
yes	1857	7mNODE_379_length_1857_cov_6.147399_component_116
yes	1857	15mNODE_423_length_1857_cov_2.832948_component_203
yes	1865	7plNODE_6481_length_1865_cov_2.766974
yes	1887	15pwNODE_5120_length_1887_cov_1.440341
yes	1944	15pwNODE_4793_length_1944_cov_2.384150
yes	1944	15plNODE_4796_length_1944_cov_1.079252
yes	1966	67pwNODE_2142_length_1966_cov_6.552474
yes	1966	67mNODE_260_length_1966_cov_16.238173_component_163

yes	1970	67plNODE_2132_length_1970_cov_4.397179
yes	2011	64plNODE_7820_length_2011_cov_6.312102
yes	2012	15plNODE_4416_length_2012_cov_4.519363
yes	2018	99mNODE_710_length_2018_cov_7.544157_component_566
yes	2041	3plNODE_7076_length_2041_cov_5.609195
yes	2071	3pw03NODE_6900_length_2071_cov_3.039095
yes	2088	67plNODE_1880_length_2088_cov_3.087710
yes	2258	67pwNODE_1570_length_2258_cov_4.465978
yes	2467	99plNODE_4660_length_2467_cov_5.311538
yes	2505	64pwNODE_5116_length_2505_cov_4.049201
yes	2608	64pwNODE_4753_length_2608_cov_5.773075
yes	2608	64mNODE_909_length_2608_cov_5.779524_component_622
yes	2769	64plNODE_4257_length_2769_cov_2.615064
yes	2875	15pwNODE_2134_length_2875_cov_19.731077
yes	2995	64plNODE_3724_length_2995_cov_4.523710
yes	3040	67plNODE_825_length_3040_cov_5.986955
yes	3325	64plNODE_3076_length_3325_cov_6.439962
yes	3436	15pwNODE_1522_length_3436_cov_3.674222
yes	3511	64pwNODE_2782_length_3511_cov_10.493499
yes	3908	64pwNODE_2346_length_3908_cov_8.964824
yes	4174	1mNODE_392_length_4174_cov_4.220904_component_329
yes	4434	64plNODE_1876_length_4434_cov_11.156257
yes	4520	1mNODE_382_length_4520_cov_9.427043_component_328
yes	5548	64pwNODE_1234_length_5548_cov_19.320790
yes	5550	1pwNODE_617_length_5550_cov_11.193067
yes	5646	64pwNODE_1182_length_5646_cov_24.234825
yes	6719	15pwNODE_548_length_6719_cov_2.532008
yes	7001	67mNODE_180_length_7001_cov_4.205121_component_156
yes	8540	78mNODE_613_length_8540_cov_2.389992_component_535
yes	11216	64mNODE_611_length_11216_cov_3.542249_component_228
yes	13611	7pwNODE_122_length_13611_cov_9.557624
no	1004	1pwNODE_24190_length_1004_cov_1.424173
no	1027	30pwNODE_19620_length_1027_cov_0.803333
no	1031	1pwNODE_22889_length_1031_cov_3.564159
no	1032	1pwNODE_22865_length_1032_cov_1.430939
no	1034	99pwNODE_34311_length_1034_cov_2.424476
no	1040	1pwNODE_22509_length_1040_cov_1.667032
no	1054	64pwNODE_33762_length_1054_cov_3.368932
no	1059	64pwNODE_33407_length_1059_cov_1.565451
no	1061	64pwNODE_33257_length_1061_cov_1.641328
no	1072	3pw03NODE_22643_length_1072_cov_1.353439
no	1081	64pwNODE_31761_length_1081_cov_2.620545
no	1083	64pwNODE_31637_length_1083_cov_1.923640
no	1084	1pwNODE_20641_length_1084_cov_0.928945
no	1086	1pwNODE_20542_length_1086_cov_1.205422
no	1093	1pwNODE_20243_length_1093_cov_2.230849
no	1093	64pwNODE_30906_length_1093_cov_2.158385
no	1097	7pwNODE_15978_length_1097_cov_7.485567
no	1101	1pwNODE_19919_length_1101_cov_1.397331
no	1117	1pwNODE_19312_length_1117_cov_1.318182
no	1117	1pwNODE_19325_length_1117_cov_0.867677
no	1123	64pwNODE_28931_length_1123_cov_1.703815
no	1152	64pwNODE_27209_length_1152_cov_2.168780
no	1156	67pwNODE_8130_length_1156_cov_1.182702

no	1175	1pwNODE_17308_length_1175_cov_6.983779
no	1180	1pwNODE_17168_length_1180_cov_1.096866
no	1195	64pwNODE_24961_length_1195_cov_2.445693
no	1254	1pwNODE_15042_length_1254_cov_2.300799
no	1260	1pwNODE_14897_length_1260_cov_1.357458
no	1269	7mNODE_462_length_1269_cov_42.211033_component_5
no	1276	1pwNODE_14488_length_1276_cov_1.277633
no	1276	1pwNODE_14492_length_1276_cov_0.992167
no	1292	1pwNODE_14112_length_1292_cov_0.924464
no	1294	1pwNODE_14057_length_1294_cov_1.092545
no	1300	99pwNODE_20080_length_1300_cov_9.886616
no	1302	64pwNODE_20482_length_1302_cov_4.082553
no	1312	7mNODE_458_length_1312_cov_4.153586_component_1
no	1313	67pwNODE_5943_length_1313_cov_1.677909
no	1321	1pwNODE_13453_length_1321_cov_1.453099
no	1343	15pwNODE_11094_length_1343_cov_1.349507
no	1347	3pw03NODE_14564_length_1347_cov_3.045902
no	1369	15pwNODE_10614_length_1369_cov_1.805153
no	1376	7mNODE_447_length_1376_cov_11.542034_component_12
no	1413	78pwNODE_18331_length_1413_cov_2.000000
no	1417	67pwNODE_4874_length_1417_cov_0.962791
no	1438	64pwNODE_16182_length_1438_cov_3.326468
no	1442	64pwNODE_16072_length_1442_cov_1.935361
no	1447	1pwNODE_11059_length_1447_cov_1.393939
no	1455	99pwNODE_15341_length_1455_cov_2.677711
no	1463	3pw03NODE_12496_length_1463_cov_1.511228
no	1479	1pwNODE_10543_length_1479_cov_1.511834
no	1485	1pwNODE_10445_length_1485_cov_1.662003
no	1498	7pwNODE_9371_length_1498_cov_4.568928
no	1519	64pwNODE_14259_length_1519_cov_3.224138
no	1535	67pwNODE_3976_length_1535_cov_1.511364
no	1538	1pwNODE_9696_length_1538_cov_1.738483
no	1591	78mNODE_812_length_1591_cov_3.066257_component_428
no	1600	1pwNODE_8897_length_1600_cov_1.082824
no	1613	1pwNODE_8703_length_1613_cov_1.538358
no	1624	30mNODE_479_length_1624_cov_13.430862_component_3
no	1679	7pwNODE_7734_length_1679_cov_2.768686
no	1771	78mNODE_793_length_1771_cov_0.696472_component_118
no	1791	15mNODE_430_length_1791_cov_2.191707_component_0
no	1814	15pwNODE_5574_length_1814_cov_2.399526
no	1831	7mNODE_384_length_1831_cov_4.980634_component_1
no	1832	64mNODE_981_length_1832_cov_5.919062_component_2
no	1833	1pwNODE_6548_length_1833_cov_2.791911
no	1857	15pwNODE_5294_length_1857_cov_0.874566
no	1923	3pw03NODE_7757_length_1923_cov_1.350780
no	1938	64pwNODE_8413_length_1938_cov_3.122032
no	1963	64mNODE_970_length_1963_cov_41.703704_component_10
no	1968	1pwNODE_5591_length_1968_cov_3.152091
no	1992	64pwNODE_7965_length_1992_cov_9.383378
no	2009	64pwNODE_7832_length_2009_cov_4.174283
no	2063	30pwNODE_3576_length_2063_cov_5.926136
no	2079	67pwNODE_1894_length_2079_cov_5.146004
no	2095	15pwNODE_4051_length_2095_cov_1.551321
no	2218	64pwNODE_6460_length_2218_cov_3.099952

no	2254	64pwNODE_6247_length_2254_cov_2.414669
no	2412	64pwNODE_5511_length_2412_cov_3.182932
no	2442	7mNODE_343_length_2442_cov_6.430238_component_8
no	2456	78mNODE_757_length_2456_cov_2.702877_component_65
no	2569	64pwNODE_4882_length_2569_cov_3.168305
no	2640	30mNODE_439_length_2640_cov_11.459610_component_30
no	2686	3plNODE_4484_length_2686_cov_2.413052
no	2757	64pwNODE_4287_length_2757_cov_4.997338
no	2811	1pwNODE_2580_length_2811_cov_3.135618
no	2894	64pwNODE_3953_length_2894_cov_54.498374
no	3064	99pwNODE_2980_length_3064_cov_2.198502
no	3122	78mNODE_728_length_3122_cov_15.366277_component_19
no	3234	7pwNODE_2254_length_3234_cov_2.972642
no	3545	30mNODE_415_length_3545_cov_3.588063_component_0
no	3605	7pwNODE_1769_length_3605_cov_8.617596
no	3644	64pwNODE_2618_length_3644_cov_3.821154
no	3760	1pwNODE_1397_length_3760_cov_4.795486
no	3765	99mNODE_643_length_3765_cov_11.120946_component_264
no	3820	99mNODE_641_length_3820_cov_13.733550_component_3
no	3837	7pwNODE_1530_length_3837_cov_6.345013
no	4010	30pwNODE_974_length_4010_cov_5.784445
no	4089	64pwNODE_2170_length_4089_cov_6.163806
no	4159	7mNODE_285_length_4159_cov_3.788938_component_0
no	4214	3pw03NODE_2201_length_4214_cov_1.381209
no	4222	78pwNODE_2320_length_4222_cov_5.756044
no	4616	7pwNODE_1024_length_4616_cov_6.093785
no	4893	7mNODE_274_length_4893_cov_2.712757_component_1
no	5096	3mNODE_564_length_5096_cov_4.330248_component_39
no	5197	15mNODE_344_length_5197_cov_2.734122_component_74
no	5400	64mNODE_784_length_5400_cov_3.644794_component_8
no	5428	30mNODE_386_length_5428_cov_4.756650_component_0
no	5642	99mNODE_606_length_5642_cov_9.506437_component_3
no	5782	7mNODE_260_length_5782_cov_4.368877_component_1
no	5910	78mNODE_646_length_5910_cov_3.265433_component_533
no	5920	67mNODE_191_length_5920_cov_10.379078_component_40
no	5978	64mNODE_762_length_5978_cov_5.425568_component_6
no	6458	1mNODE_358_length_6458_cov_5.761333_component_95
no	6891	64mNODE_741_length_6891_cov_2.374335_component_577
no	7371	3mNODE_531_length_7371_cov_4.048178_component_0
no	7560	64pwNODE_672_length_7560_cov_2.060406
no	8068	64mNODE_720_length_8068_cov_2.476137_component_228
no	9060	67mNODE_171_length_9060_cov_6.221314_component_6
no	9154	78mNODE_608_length_9154_cov_9.694029_component_395
no	9432	3mNODE_506_length_9432_cov_4.282644_component_379
no	9478	64mNODE_700_length_9478_cov_6.132927_component_222
no	9912	30mNODE_322_length_9912_cov_4.737251_component_0
no	9980	64mNODE_691_length_9980_cov_5.305998_component_222
no	10373	1mNODE_316_length_10373_cov_3.381612_component_312
no	10520	15mNODE_284_length_10520_cov_2.872222_component_1
no	10916	3mNODE_467_length_10916_cov_2.565669_component_0
no	10998	99mNODE_500_length_10998_cov_8.281391_component_374
no	11003	64mNODE_623_length_11003_cov_2.801949_component_548
no	11089	7mNODE_187_length_11089_cov_4.009943_component_0
no	11177	7mNODE_182_length_11177_cov_5.770045_component_63

no	11226	78pwNODE_431_length_11226_cov_2.839625
no	11811	99pwNODE_236_length_11811_cov_3.059141
no	11861	67mNODE_131_length_11861_cov_6.132947_component_4
no	11867	15mNODE_266_length_11867_cov_1.359114_component_178
no	11980	99mNODE_444_length_11980_cov_3.287100_component_441
no	12023	64mNODE_556_length_12023_cov_4.829186_component_2
no	12099	78mNODE_491_length_12099_cov_3.906674_component_7
no	12676	7mNODE_152_length_12676_cov_4.791298_component_13
no	13066	7mNODE_147_length_13066_cov_9.176598_component_36
no	13710	7mNODE_138_length_13710_cov_3.990429_component_12
no	13923	99mNODE_359_length_13923_cov_3.897724_component_357
no	15396	30mNODE_238_length_15396_cov_3.965158_component_30
no	15575	7mNODE_118_length_15575_cov_3.200155_component_56
no	15725	78mNODE_346_length_15725_cov_6.408450_component_34
no	16084	78mNODE_340_length_16084_cov_2.379144_component_308
no	16490	78mNODE_331_length_16490_cov_2.802787_component_299
no	16567	99pwNODE_100_length_16567_cov_2.851764
no	16890	78mNODE_321_length_16890_cov_1.962179_component_98
no	17369	78mNODE_312_length_17369_cov_2.872231_component_283
no	19359	99mNODE_179_length_19359_cov_2.296433_component_184
no	19666	99mNODE_174_length_19666_cov_2.701929_component_179

Anexo 2. *Contigs* plasmidiais recuperados com similares no RefSeq

% IDENTITY	LENGHT QUERY	LENGHT REFERENCE	COVERAGE QUERY	COVERAGE REFERENCE	*CONTIG*	REFERENCE
99.93	3436	3279	43.71	45.78	15pwNODE_1522_length_3436_cov_3.674222	NC_022993.1-Geobacillus-sp.-610-plasmid-pGTD7-
99.54	3436	3279	50.70	53.10	15pwNODE_1522_length_3436_cov_3.674222	NC_022993.1-Geobacillus-sp.-610-plasmid-pGTD7-
97.82	2218	2427	16.50	15.12	64pwNODE_6460_length_2218_cov_3.099952	NZ_CP042565.1-Acinetobacter-baumannii-strain-E47-plasmid-pE47_009-
91.90	2218	2427	32.28	29.50	64pwNODE_6460_length_2218_cov_3.099952	NZ_CP042565.1-Acinetobacter-baumannii-strain-E47-plasmid-pE47_009-
95.34	2218	2427	51.26	46.85	64pwNODE_6460_length_2218_cov_3.099952	NZ_CP042565.1-Acinetobacter-baumannii-strain-E47-plasmid-pE47_009-
88.43	1938	2427	20.07	16.03	64pwNODE_8413_length_1938_cov_3.122032	NZ_CP042565.1-Acinetobacter-baumannii-strain-E47-plasmid-pE47_009-
100.00	1992	4136	66.82	32.18	64pwNODE_7965_length_1992_cov_9.383378	NZ_LN873255.1-Acinetobacter-johnsonii-strain-LS47-1-plasmid-pAJOLS1.1-
97.37	1992	4135	66.82	32.16	64pwNODE_7965_length_1992_cov_9.383378	NZ_LN873256.1-Acinetobacter-lwoffii-strain-ED23-35-plasmid-pALWED1.8-
100.00	1992	4136	33.18	15.98	64pwNODE_7965_length_1992_cov_9.383378	NZ_LN873255.1-Acinetobacter-johnsonii-strain-LS47-1-plasmid-pAJOLS1.1-
99.58	1535	4135	46.25	17.15	67pwNODE_3976_length_1535_cov_1.511364	NZ_LN873256.1-Acinetobacter-lwoffii-strain-ED23-35-plasmid-pALWED1.8-
99.03	1535	4136	53.75	19.95	67pwNODE_3976_length_1535_cov_1.511364	NZ_LN873255.1-Acinetobacter-johnsonii-strain-LS47-1-plasmid-pAJOLS1.1-
96.97	1535	4135	53.55	19.90	67pwNODE_3976_length_1535_cov_1.511364	NZ_LN873256.1-Acinetobacter-lwoffii-strain-ED23-35-plasmid-pALWED1.8-