

Universidade Tecnológica Federal do Paraná
Curso de Engenharia Eletrônica

Roberto Marafon Leandro

**Análise de dados de metagenômica obtidos de
amostras de solo: Caracterização taxonômica,
do resistoma e do viruloma**

Toledo
2025

Roberto Marafon Leandro

Análise de dados de metagenômica obtidos de amostras de solo: Caracterização taxonômica, do resistoma e do viruloma

**Analysis of metagenomics data obtained from soil samples:
Taxonomic, resistome, and virulome characterization**

Trabalho de Conclusão de Curso apresentado ao Curso de Engenharia Eletrônica da Universidade Tecnológica Federal do Paraná, Campus Toledo, como requisito parcial para a obtenção do título de Bacharel em Engenharia Eletrônica.

Orientador Eduardo Vinícius Kuhn
Coorientadora Glacy Jaqueline da Silva
Coorientador Emerson Willian Danzer

Toledo
2025

Esta licença permite que outros remixem, adaptem e criem a partir do trabalho licenciado para fins não comerciais, com crédito atribuído ao autor. Os usuários não têm que licenciar os trabalhos derivados sob os mesmos termos estabelecidos pelo autor do trabalho original.



4.0 Internacional

AGRADECIMENTOS

Registro aqui os meus sinceros agradecimentos a todos que, de alguma forma, contribuíram para o desenvolvimento deste trabalho. Em especial:

Agradeço à minha família, minha mãe Maristela Marafon Leandro, meu pai Diomar Leandro, meu irmão Julio Marafon Leandro e ao namorado Eduardo Henrique Fiametti, pelo suporte, pelo incentivo aos estudos e por todo o amor direcionado a mim.

Agradeço aos meus amigos pela companhia e pelos conselhos.

Agradeço ao meu orientador Eduardo Vinícius Kuhn, e aos meus coorientadores Glacy Jaqueline da Silva e Emerson William Danzer, pelo tempo, atenção, orientação e ensinamentos.

Agradeço ao professor Sérgio Paulo Dejato da Rocha, e seus alunos Arthur Bossi do Nascimento e Bruno Henrique Dias de Oliva, assim como ao professor Henrique dos Santos Felipetto pelos auxílios prestados durante o desenvolvimento deste trabalho.

Agradeço ao Fábio Bays de Araujo pelos *scripts* de importação dos dados taxonômicos gerados pelo *software Kraken2*.

Agradeço aos membros da banca examinadora, Jefferson Gustavo Martins e Evandro Marcos Kolling, pelas pertinentes considerações, contribuições e sugestões.

Agradeço à Universidade Tecnológica Federal do Paraná, em especial à Coordenadoria de Gestão de Tecnologia da Informação (COGETI), pela infraestrutura disponibilizada.

Dedico este trabalho à minha mãe Maristela
e ao meu falecido pai Diomar.

RESUMO

Nas últimas décadas, diversos avanços tecnológicos e incentivos voltados à pesquisa científica vêm fazendo com que o Brasil obtenha destaque internacional no âmbito da produção agrícola frente ao aumento da demanda global por alimentos. Apesar disso, verifica-se ainda significativa negligência em relação à importância das comunidades microbianas, seja por falta de informação ou pela subestimação de seu papel crucial no processo, o que resulta em práticas de manejo que não conseguem explorar plenamente a capacidade produtiva do solo. Para mudar esse cenário, uma abordagem promissora se dá através do uso de técnicas avançadas de biotecnologia, tal como o sequenciamento metagenômico de amostras de solo. Análises desses dados oriundos do sequenciamento podem revelar informações valiosas sobre os microrganismos presentes na amostra, permitindo assim uma gestão mais eficiente e sustentável dos solos agrícolas. Nesse contexto, busca-se aqui conduzir análises de dados de metagenômica obtidos de amostras de solo, utilizando-se de dados (brutos) sequenciados disponibilizados publicamente por Ordine et al. (2023). Especificamente, tem-se por objetivo i) caracterizar o perfil (taxonômico) das comunidades microbianas, evidenciando a diversidade e a abundância de microrganismos; ii) identificar o perfil do resistoma e do viruloma, focando na abundância e diversidade dos genes de resistência a antibióticos (ARGs) e de fatores de virulência (VFs); bem como iii) clarificar as diferenças observadas nas amostras de solo analisadas, levando em consideração as características de cada amostra. Para tal, diferentes ferramentas de bioinformática são aplicadas em sequência a fim de produzir *pipelines* capazes de realizar a análise de dados de forma não assistida. Em particular, duas *pipelines* são implementadas aqui utilizando linguagem GNU Bash, sendo uma delas destinada à classificação taxonômica e a outra para a caracterização de genes de resistência a antibióticos (ARGs) e de fatores de virulência (VFs). Como resultado da execução dessas pipelines, gêneros importantes para a saúde do solo e das plantas foram identificados, tais como *Bradyrhizobium*, *Nocardioides* e *Pseudomonas*. Por sua vez, com respeito dos ARGs, genes relevantes como *vanRO*, *rsmA* e *RbpA* foram identificados. Ainda, no contexto de VFs, a presença de genes como *acpXL*, *hsbB1* e *pilG* foram observados nas amostras analisadas. Portanto, os resultados obtidos ratificam a validade das *pipelines* implementadas tanto no que tange à classificação taxonômica quanto à caracterização do resistoma e do viruloma.

Palavras-chave: Bioinformática; Comunidades microbianas; Resistência antimicrobiana; Saúde do solo.

ABSTRACT

In recent decades, various technological advances and incentives directed towards scientific research have made Brazil stand out internationally in the field of agricultural production, in response to the growing global demand for food. Despite this, there is still significant neglect regarding the importance of soil microbial communities, whether due to a lack of information or underestimation of their crucial role, which results in management practices that fail to fully exploit the productive capacity of the soil. To change this scenario, a promising approach is through the use of advanced biotechnology techniques, such as metagenomic sequencing of soil samples. Analyses of data from this sequencing can reveal valuable information about the microorganisms present in the sample, thus enabling more efficient and sustainable management of agricultural soils. In this context, the aim here is to proceed with metagenomic data analyses from soil samples, using (raw) sequenced data made publicly available by Ordine et al. (2023). Specifically, the objectives are to: i) characterize the taxonomic profile of microbial communities, highlighting the diversity and abundance of microorganisms; ii) identify the resistome and virulome profiles, focusing on the abundance and diversity of antibiotic resistance genes (ARGs) and virulence factors (VFs); and iii) clarify the observed differences in the analyzed soil samples, considering the adopted management practices. To this end, various existing bioinformatics tools are applied in sequence to implement pipelines capable of performing unsupervised data analysis. In particular, this work implements two pipelines using the GNU Bash programming language, with one of them being designed for taxonomic classification and the other for characterizing antibiotic resistance genes (ARGs) and virulence factors (VFs). As a result of the execution of these pipelines, important genera for soil and plant health were identified, such as *Bradyrhizobium*, *Nocardioides*, and *Pseudomonas*. In turn, regarding the ARGs, relevant genes such as *vanRO*, *rsmA*, and *RbpA* were identified. Still, in the context of VFs, the presence of genes such as *acpXL*, *hsbB1*, and *pilG* were observed in the samples considered. Therefore, the results obtained confirm the validity of the implemented pipelines, which were designed for both taxonomic classification and characterization of the resistome and the virulome.

Keywords: Bioinformatics; Microbial communities; Antimicrobial resistance; Soil health.

LISTA DE ILUSTRAÇÕES

Figura 1 – Localização geográfica, no estado de São Paulo, dos pontos/sítios de coleta das amostras de solo consideradas [conforme (ORDINE et al., 2023, Table 1)].	24
Figura 2 – Diagrama de blocos ilustrando as <i>pipelines</i> desenvolvidas. (a) Etapas exclusivas à <i>pipeline</i> de classificação taxonômica. (b) Etapas exclusivas à <i>pipeline</i> de caracterização do resistoma e do viruloma.	27
Figura 3 – Qualidade média das bases após a etapa de controle de qualidade e limpeza para a <i>forward read</i> (linha escura) e a <i>reverse read</i> (linha clara). (a) SRR22278225. (b) SRR22278226. (c) SRR22278227. (d) SRR22278228. (e) SRR22278229. (f) SRR22278230. (g) SRR22278231. (h) SRR22278232.	45
Figura 4 – Análise taxonômica considerando a abundância relativa dos gêneros presentes na amostra (barras cinzas) e a média do conjunto (linha escura sólida). (a) SRR22278225. (b) SRR22278226. (c) SRR22278227. (d) SRR22278228. (e) SRR22278229. (f) SRR22278230. (g) SRR22278231. (h) SRR22278232.	46
Figura 5 – Genes de resistência a antibióticos identificados nas diferentes amostras analisadas.	48
Figura 6 – Genes de fatores de virulência identificados nas diferentes amostras analisadas.	49
Figura 7 – Fatores de virulência relacionados aos genes identificados nas diferentes amostras.	49
Figura 8 – Funções associadas aos genes de fatores de virulência identificados nas diferentes amostras analisadas.	50
Figura 9 – Cepas bacterianas cujo material genético foi alinhado aos genes de fatores de virulência identificados nas diferentes amostras analisadas. . . .	50
Figura 10 – Comparação entre os genes encontrados aqui e aqueles de Ordine et al. (2023) (identificados pelo símbolo '*'). (a) Genes de resistência a antibióticos (ARGs). (b) Genes de fatores de virulência (VFs).	63

LISTA DE TABELAS

Tabela 1 – Principais informações relacionadas às amostras de solo consideradas	25
Tabela 2 – Principais métricas das amostras antes e depois da etapa de controle de qualidade e limpeza.	43
Tabela 3 – Características dos <i>contigs</i> montados pelo <i>software</i> MEGAHIT.	44
Tabela 4 – Resultados decorrentes da etapa de predição de genes.	47

LISTA DE SCRIPTS

<i>Script 1</i> – Obtenção dos dados de metagenômica (amostras) do SRA.	28
<i>Script 2</i> – Controle de qualidade e limpeza dos dados de metagenômica.	29
<i>Script 3</i> – Classificação taxonômica sobre os dados filtrados.	30
<i>Script 4</i> – Montagem dos <i>contigs</i> pelo <i>software MEGAHIT</i>	31
<i>Script 5</i> – Predição de genes a partir dos <i>contigs</i> montados.	32
<i>Script 6</i> – Identificação de genes de interesse a partir de dados anotados.	33
<i>Script 7</i> – Criação do arquivo .csv contendo todos os níveis taxonômicos no banco de dados utilizado pelo Kraken2	35
<i>Script 8</i> – Criação do arquivo .csv contendo os níveis taxonômicos e as correspondentes abundâncias de cada amostra.	36
<i>Script 9</i> – Código responsável por evocar as funções implementadas para gerar o arquivo .csv com as informações de taxonomia das diferentes amostras.	37
<i>Script 10</i> – Trecho de código que evoca as funções responsáveis por gerar os gráficos de barras dos gêneros taxonômicos caracterizados.	37
<i>Script 11</i> – Definição da função responsável por gerar os gráficos de barras dos gêneros taxonômicos caracterizados.	38
<i>Script 12</i> – Trecho de código responsável por evocar as funções de importação dos dados e de geração de figuras acerca dos genes identificados.	39
<i>Script 13</i> – Definição da função responsável pela importação e formatação dos arquivos .tsv gerados durante a etapa de identificação de genes.	40
<i>Script 14</i> – Definição da função de criação do <i>dataframe</i> contendo a informação gênica de todos os arquivos importados.	40
<i>Script 15</i> – Função responsável por gerar as figuras ilustrando os ARGs e os VFs identificados.	41

SUMÁRIO

1	INTRODUÇÃO	11
1.1	Objetivos	13
1.2	Organização do documento	14
2	MATERIAIS E MÉTODOS	15
2.1	Fluxo de trabalho envolvendo metagenômica	15
2.2	Revisitando etapas e ferramentas de bioinformática	18
2.2.1	Limpeza e controle de qualidade	19
2.2.2	Classificação taxonômica	19
2.2.3	Montagem metagenômica	20
2.2.4	Predição de genes	22
2.2.5	Identificação de genes	22
2.3	Descrição do conjunto de dados	23
3	<i>PIPELINES PARA PROCESSAMENTO E ANÁLISE DE DADOS</i>	26
3.1	Obtenção dos dados de metagenômica (amostras)	26
3.2	Controle de qualidade e limpeza	27
3.3	Classificação taxonômica	28
3.4	Montagem de <i>contigs</i>	30
3.5	Predição de genes	31
3.6	Identificação de genes de interesse	32
3.7	Manipulação dos arquivos gerados pelo Kraken2	34
3.8	Manipulação dos arquivos gerados pelo ABRicate	34
4	RESULTADOS E DISCUSSÕES	42
4.1	Filtragem dos dados	42
4.2	Análise taxonômica	43
4.3	Estatísticas relativas à montagem de <i>contigs</i>	44
4.4	Resultados decorrentes da predição de genes	47
4.5	Caracterização do resistoma	47
4.6	Caracterização do viruloma	48
5	CONSIDERAÇÕES FINAIS	51
5.1	Conclusões	51
5.2	Sugestões de trabalhos futuros	51
5.3	Trabalhos publicados	52

REFERÊNCIAS 53

APÊNDICE – COMPARAÇÕES COM ORDINE ET AL. (2023) 62

1 INTRODUÇÃO

Na produção agrícola, o Brasil figura dentre os 10 maiores produtores mundiais de grãos, assumindo posições de liderança na exportação de diversas *commodities*. Especificamente em 2020, o país se destacou como o 4º maior produtor de grãos (239 milhões de toneladas) e o 2º maior exportador (123 milhões de toneladas), o que corresponde a 19% do total de grãos exportados no mundo e representa um montante de 37 bilhões de dólares (ARAGÃO; CONTINI, 2023). Dentre os principais grãos, o Brasil figura como o maior produtor e exportador de soja, desde a safra 2019/2020; em especial, na safra de 2022/2023, o país cultivou 42% de todo o soja produzido no mundo, além de ter participado com cerca de 27% de todo o milho exportado globalmente (maior exportador desse grão) (AGROADVANCE, 2024). Bens e serviços atrelados ao agronegócio brasileiro movimentaram em torno de R\$ 2,54 trilhões em 2022, o que equivale a 25% do PIB; desse montante, R\$ 1,836 trilhão diz respeito exclusivamente ao ramo agrícola (CNA, 2024). O agronegócio brasileiro desempenha um papel fundamental não apenas para a economia nacional, mas também na segurança alimentar global e na promoção da saúde humana já que, através da produção e exportação de alimentos essenciais para dietas balanceadas e saudáveis, o país contribui sobremaneira para a redução da fome e desnutrição em diversas regiões do mundo (FAO, IFAD, UNICEF, WFP e WHO, 2024). Parte dessa trajetória de sucesso da agricultura brasileira vêm sendo construída e sustentada, ao longo das últimas décadas, por uma série de políticas públicas¹ assim como por investimentos significativos em pesquisa, desenvolvimento e inovação (PD&I) voltadas ao aprimoramento e consolidação das técnicas de manejo.

Com respeito às técnicas de manejo adotadas na produção agrícola, destacam-se o plantio direto, o uso de agroquímicos (fertilizantes e agrotóxicos), a correção físico-química do solo, o monitoramento de indicadores biológicos (e.g., respiração basal, biomassa microbiana, atividade enzimática, comprimento das hifas e coeficiente metabólico), bem como o melhoramento genético de sementes. Essas técnicas visam, sobretudo, maximizar a produtividade, assegurar o controle de pragas e doenças, promover o uso sustentável do solo, melhorar a disponibilidade de nutrientes e aumentar a resistência a condições climáticas

¹ Dentre as políticas públicas determinantes para posicionar o Brasil como um dos mais importantes produtores no cenário internacional, destacam-se a criação i) da EMBRAPA (Empresa Brasileira de Pesquisa Agropecuária) em 1973, visando o desenvolvimento de tecnologias adaptadas às condições tropicais brasileiras; ii) de incentivos ao Crédito Rural subsidiado para facilitar o acesso de agricultores ao financiamento necessário para modernização de suas operações (e.g., insumos, máquinas); iii) do Programa Nacional de Irrigação em 1979 a fim de aumentar a produtividade por permitir a produção durante o ano todo, independentemente das condições climáticas; iv) da Política de Preços Mínimos para certos produtos, assegurando que os agricultores tivessem uma renda mínima garantida e incentivando a produção; v) de investimentos em infraestrutura de transporte, visando facilitar o escoamento da produção agrícola; vi) de Programas de Capacitação e Extensão Rural para promover a adoção de novas tecnologias e práticas agrícolas mais eficientes (EMBRAPA, 2022).

adversas (FAO, 2017; BUENO et al., 2021; FUENTES-LLANILLO et al., 2021). Contudo, apesar de representarem avanços importantes, tais técnicas de manejo não exploram plenamente a capacidade produtiva do solo por negligenciar características importantes do microbioma edáfico² (WILHELM et al., 2023; BANERJEE; HEIJDEN, 2023; COMPANT et al., 2024). Vale lembrar que os microrganismos presentes no solo contribuem significativamente para a manutenção da saúde das plantas, visto que participam de atividades envolvendo a ciclagem de nutrientes, estrutura do solo, transformações de carbono, crescimento, regulação hormonal, controle de estresse, resistência sistêmica a patógenos, absorção de água, mineralização, dentre outros (BANERJEE; HEIJDEN, 2023; COMPANT et al., 2024). Ademais, além de não explorarem plenamente a capacidade do solo, as técnicas de manejo adotadas por vezes alteram drasticamente o microbioma edáfico, devido ao uso excessivo de agroquímicos nas plantações, à aplicação de dejetos/esterco enriquecido com antibióticos (proveniente da produção pecuária), e ao aumento do desmatamento para fins agrícolas ou urbanos (ORDINE et al., 2023). Essas práticas modificam o resistoma³ e o viruloma⁴ intrínseco do solo (i.e., alteram as comunidades bacterianas), favorecendo a disseminação de genes de resistência a antibióticos (ARGs) e de fatores de virulência (VFs). Como consequência, tem-se a predominância de organismos com resistência adquirida e maior potencial patogênico, decorrente do aumento na taxa de transferência horizontal de ARGs e VFs, intensificando assim a resistência antimicrobiana (AMR) no ambiente (ORDINE et al., 2023). Vale destacar que, recentemente, as Nações Unidas reconheceram os possíveis perigos causados pela falta de vigilância e descontrole sobre a AMR, sendo esse um problema a ser atacado de forma coordenada globalmente (MAPA, 2024). Portanto, o desenvolvimento, aperfeiçoamento e adoção de novas tecnologias que levam em conta o microbioma edáfico são fundamentais para salvaguardar a saúde e aprimorar a eficiência do solo assim como para preservar a saúde humana, possibilitando uma gestão mais eficiente, competitiva e sustentável dos setores agrícola brasileiro.

Uma abordagem promissora se dá através do uso de técnicas avançadas de biotecnologia, tal como o sequenciamento por metagenômica (*metagenomics sequencing*) de amostras de solo. Essa abordagem (em especial, o sequenciamento *shotgun*) vêm ganhando cada vez mais espaço devido à capacidade de sequenciar toda a informação genética em uma dada amostra analisada, independentemente da diversidade de organismos presentes e sem a necessidade de isolamento ou cultivo específico (QUINCE et al., 2017; WARNECKE et al., 2007; BHARTI; GRIMM, 2021). Dessa forma, torna-se possível obter informações genéticas mais abrangentes acerca do microbioma edáfico presente em terras agricultáveis, em contraste com a abordagem de sequenciamento de *amplicons*⁵ que é

² Comunidade de microrganismos, incluindo bactérias, fungos, arqueias e vírus, que habitam o solo.

³ Conjunto de genes de resistência a antibióticos presentes em um dado microbioma, cuja compreensão é fundamental para estudar a disseminação da resistência a antibióticos e o impacto na saúde pública.

⁴ Conjunto de genes relacionados à virulência de um determinado microbioma (SALGUEIRO et al., 2023).

⁵ Vale mencionar que o sequenciamento de *amplicons*, também conhecido como sequenciamento

voltada à identificação de um único gene ou regiões-alvo específicas do DNA necessárias para caracterizar espécies de fungos e bactérias (BHARTI; GRIMM, 2021; LOBANOV; GOBET; JOYCE, 2022). Análises dos dados oriundos do sequenciamento podem revelar informações valiosas sobre os microrganismos presentes na amostra, possibilitando assim caracterizar o perfil taxonômico (abundância e diversidade) das comunidades microbianas, identificar genes relacionados à funções biológicas (e.g., fixação de nitrogênio e carbono, solubilização de fosfato, produção de hormônios vegetais, matéria orgânica) e verificar o perfil do resistoma e do viruloma a partir da presença de ARGs e VFs. Contudo, tais análises trazem importantes desafios do ponto de vista prático (sobretudo, quanto à demanda por recursos computacionais), os quais decorrem do grande volume de dados gerado a partir do sequenciamento. Ademais, tendo em vista a diversidade de ferramentas de bioinformática disponíveis bem como as suas correspondentes vantagens e limitações, a escolha da ferramenta mais adequadas em cada uma das etapas de análise se torna uma tarefa complexa (BHARTI; GRIMM, 2021; LIU et al., 2021). Ainda, vale destacar que não existem até então *pipelines*⁶ consolidadas para conduzir a análise de dados de metagenômica devido à rápida evolução das tecnologias envolvidas, tornando obsoletas até mesmo aquelas *pipelines* populares de laboratórios internacionais (RICHARDSON et al., 2022).

1.1 OBJETIVOS

Neste contexto, o presente trabalho busca conduzir análises de dados de metagenômica obtidos de amostras de solo, focando sobre os dados (brutos) sequenciados disponibilizados publicamente por Ordine et al. (2023). Especificamente, utilizando-se de ferramentas de bioinformática, tem-se aqui por objetivo

- determinar o perfil (taxonômico) das comunidades microbianas, evidenciando a diversidade e a abundância de microrganismos;
- caracterizar o perfil do resistoma e do viruloma, focando na abundância e diversidade de ARGs e de VFs presentes na amostra; bem como
- clarificar as diferenças observadas nas amostras analisadas, levando em conta as suas características.

Vale salientar que este trabalho visa, sobretudo, contribuir com a construção de *pipelines* para a análise de dados de metagenômica.

^{16S/18S/ITS, metabarcoding, metataxonomics ou community profiling, visa rastrear variantes ou organismos específicos, sendo comumente utilizado para identificar espécies individuais em culturas puras e detectar organismos de interesse (e.g., patógenos).}

⁶ O termo se refere a um programa, um *script* ou a uma ordem específica de programas em que um certo conjunto de operações deve ser executado (LIU et al., 2021).

1.2 ORGANIZAÇÃO DO DOCUMENTO

O presente trabalho está organizado como segue. A Seção 2 aborda um fluxo típico de trabalho de um projeto de metagenômica do tipo *shotgun*, descrevendo as diferentes etapas desde a coleta das amostras até as análises de bioinformática, assim como as principais operações e *softwares* empregados durante as análises de bioinformática. A Seção 3 descreve as *pipelines* desenvolvidas aqui para processar os dados de metagenômica relativos às amostras, bem como os *scripts* implementados para conduzir a análise dos resultados. A Seção 4 apresenta os resultados intermediários e finais obtidos, discutindo tanto sobre as estatísticas referentes às etapas de controle de qualidade e limpeza, de montagem e de predição de genes, quanto sobre o perfil taxonômico, do resistoma e do viruloma. Por fim, a Seção 5 traz algumas das considerações finais deste trabalho, evidenciando as principais conclusões alcançadas, as sugestões de trabalhos futuros e a lista de trabalho publicados e/ou em vias de publicação.

2 MATERIAIS E MÉTODOS

Nesta seção, alguns conceitos fundamentais necessários para o desenvolvimento do presente trabalho são revisitados. Primeiramente, o fluxo de trabalho típico envolvendo estudos de metagenômica é descrito, visando fornecer um entendimento das diferentes etapas envolvidas desde a coleta e preparo do DNA até a obtenção dos dados brutos sequenciados e análises de bioinformática. Em seguida, algumas etapas e ferramentas fundamentais para a análise dos dados de metagenômica obtidos do sequenciamento são discutidas, destacando sobretudo aquelas etapas e ferramentas consideradas aqui. Por fim, o conjunto de dados utilizado é apresentado, contendo dados brutos advindos do sequenciamento por metagenômica de amostras de solo juntamente com algumas informações associadas às especificidades dos diferentes sítios de coleta.

2.1 FLUXO DE TRABALHO ENVOLVENDO METAGENÔMICA

Tipicamente, um fluxo de trabalho de um projeto envolvendo metagenômica *shotgun*¹ segue as mesmas 6 etapas, a saber:

- 1) **Preparação da amostra:** Essa etapa tem início com a coleta cuidadosa de uma amostra de materiais do ambiente-alvo (e.g., solo, água, fezes, *swabs*, tecidos), utilizando técnicas que visem minimizar a contaminação e garantir representatividade adequada. Tal amostra, após coletada, deve ser adequadamente conservada (geralmente, em condições de baixa temperatura) durante o transporte e/ou armazenamento para preservar a integridade do material genético. É fundamental evitar contaminações com DNA estranho ou genoma do hospedeiro², especialmente durante o manuseio. Portanto, o método de coleta, transporte e armazenamento da amostra, bem como as características do ambiente-alvo introduzem variáveis que precisam ser consideradas e devidamente documentadas (na forma de *metadados*) para assegurar a confiabilidade dos resultados (QUINCE et al., 2017; BHARTI; GRIMM, 2021).
- 2) **Extração de DNA:** Essa etapa, necessária especialmente no sequenciamento do tipo *shotgun*, consiste da separação do DNA da amostra de outras impurezas e contaminantes. Isso envolve a lise celular (ruptura da membrana), utilizando mé-

¹ Em contraste com o sequenciamento metagenômico de *amplicons* 16S/18S/ITS que foca no sequenciamento de regiões-alvo específicas do DNA, o sequenciamento metagenômico *shotgun* fornece informações genéticas sobre todos os genes dos organismos presentes em uma amostra; em outras palavras, o DNA total da amostra é sequenciado utilizando a tecnologia de sequenciamento de nova geração (NGS), evitando assim a necessidade de isolamento e cultivo de microrganismos.

² Caso o genoma do hospedeiro seja conhecido, é possível removê-lo dos dados obtidos do sequenciamento.

todos mecânicos, químicos ou enzimáticos (BHARTI; GRIMM, 2021), para liberar o conteúdo interno da célula que inclui o DNA. Após a lise celular, o DNA é então coletado por um processo de precipitação por centrifugação, enquanto o sobrenadante contendo impurezas e outros materiais contaminantes (e.g., RNAs, proteínas e outros compostos celulares) é descartado (QUINCE et al., 2017; QIAGEN, 2024). Dependendo das particularidades do ambiente-alvo, protocolos usando *kits* personalizados são adotados para otimizar a extração (e.g., o *DNeasy PowerSoil Pro Kit*³ em amostras de solo e lodo, o *DNeasy PowerWater Kit* em amostras de água, *QIAamp Fast DNA Stool Mini Kit* da QIAGEN em amostras fecais complexas e ricas em matéria orgânica). Logo, é evidente que a adoção de um dado protocolo de extração de DNA deve levar em conta o ambiente-alvo em questão (QIAGEN, 2024).

- 3) **Quantificação e controle de qualidade:** Essa etapa visa quantificar a concentração de DNA presente na amostra e avaliar sua adequação dentro de critérios de qualidade usuais. Em particular, a quantificação é comumente realizada por 3 métodos principais, a saber: i) a espectrofotometria, usando o equipamento Nano-Drop da ThermoFisher devido à sua simplicidade, necessidade de apenas 1 µl da amostra e precisão; ii) a fluorometria, sendo o equipamento Qubit da ThermoFisher especialmente útil (devido a sensibilidade e especificidade) para quantificar baixas concentrações de DNA em amostras já que mede a fluorescência de corantes que se ligam ao DNA (logo, não sofre interferência de contaminantes); e iii) a eletroforese em gel, que separa as moléculas de DNA pelo tamanho, permitindo a visualização das bandas de DNA em gel de agarose e a comparação com padrões para estimar a concentração de DNA. Por sua vez, o controle de qualidade foca em garantir que o DNA extraído esteja livre de contaminantes e/ou degradação. Técnicas como a medição da relação A260/A280 (para pureza) e a análise de integridade em gel de agarose são comumente empregadas. Dessa forma, garante-se que o DNA esteja em condições apropriadas, tanto em termos de integridade quanto de pureza, para dar continuidade ao processo com a construção das bibliotecas de sequenciamento (GREEN; SAMBROOK, 2012).
- 4) **Construção da biblioteca:** Essa etapa trata da fragmentação das fitas de DNA em tamanhos menores, geralmente entre 250 e 300 bp (*base pair*), utilizando métodos mecânicos, enzimáticos ou químicos. Após a fragmentação, sequências específicas de nucleotídeos, denominadas adaptadores, são ligados às extremidades dos fragmentos, atuando como códigos de barras únicos que permitem a identificação precisa do início/fim de cada fragmento durante o sequenciamento. Uma seleção de tamanho é frequentemente realizada para padronizar a biblioteca e remover adaptadores livres, o que se dá por meio da adição de esferas magnéticas que possuem afinidade

³ O protocolo inclui a quebra mecânica das células microbianas, seguida de etapas de purificação para remover impurezas.

seletiva para fragmentos de DNA de determinado tamanho. Isso assegura que os fragmentos estejam em condições ideais para a plataforma de sequenciamento. Esses cuidados durante a construção das bibliotecas, visam maximizar a precisão/eficiência do sequenciamento e minimiza potenciais impactos na representatividade dos dados (VAN DIJK; JASZCZYSZYN; THERMES, 2014; ILLUMINA, 2024a).

- 5) **Sequenciamento:** Essa etapa consiste na amplificação e leitura dos diversos fragmentos de DNA presentes na amostra e, posterior, conversão dessas leituras para uma representação em ambiente computacional. Para isso, tecnologias de sequenciamento de nova geração (NGS)⁴, como o método de sequenciamento por síntese (SBS) da Illumina, ganharam notoriedade por sua capacidade de sequenciamento simultâneo (do tipo *shotgun*) de milhões de fragmentos em um curto período de tempo⁵ (sem a necessidade de isolamento e cultivo de microrganismos) (HU et al., 2021; ILLUMINA, 2024d). Nesse método SBS, nucleotídeos fluorescentes são incorporados às fitas de DNA e as informações genéticas são capturadas por leituras ópticas, as quais são inicialmente armazenadas em arquivos com extensão `.bcl`. Em seguida, essas leituras brutas são demultiplexadas, i.e., separadas com base nos adaptadores específicos. Por fim, as sequências de nucleotídeos demultiplexadas são armazenadas em arquivos com extensão `.fastq`⁶ relativos a amostra considerada, os quais são utilizados em análises subsequentes (ILLUMINA, 2024c). Decorrente dessa abordagem de sequenciamento metagenômico *shotgun*, grandes volumes de dados são gerados que carecem de análises de bioinformática adequadas (BHARTI; GRIMM, 2021).
- 6) **Análises bioinformática** (*foco do presente trabalho*): Essa etapa, crucial para extrair informações úteis a partir dos dados obtidos do sequenciamento (XU et al., 2014), envolve a escolha e execução de diferentes ferramentas de bioinformática organizadas em uma *pipeline* (conforme detalhado na Seção 2.2). Tais *pipelines* visam (QUINCE et al., 2017)
 - revelar as composições (abundância e diversidade) taxonômica e funcional das comunidades microbianas na amostra, possibilitando identificar quais microrganismos estão presentes assim como inferir sobre o papel que desempenham no ambiente em que estão inseridos;

⁴ Essa tecnologia revolucionou a pesquisa em biotecnologia, permitindo estudos inéditos sobre sistemas biológicos, como o sequenciamento rápido de genomas inteiros, a análise de regiões nucleotídicas específicas, fatores epigenéticos e o microbioma humano (ILLUMINA, 2024e).

⁵ Vale destacar que equipamentos modernos, como o *NovaSeq 6000* da Illumina, são capazes de processar até 6000 Gb em uma única corrida, com leituras curtas (*short reads*) de até 250 bp por fita e tempos de operação variando entre 13 e 44h.

⁶ O formato `.fastq` vem sendo adotado como padrão para os dados gerados como saída pela maioria das plataformas de sequenciamento (CHEN, 2023); para detalhes sobre o formato, veja (ILLUMINA, 2024b).

- construir catálogos gênicos que são fundamentais para compreender a diversidade genética, a abundância de genes, as possíveis funções biológicas (envolvidas em processos vitais como a degradação de matéria orgânica, ciclagem de nutrientes e resistência ao estresse), genes específicos (tais como aqueles associados à resistência a pesticidas, ARGs e VFs, degradação de poluentes ambientais e biossíntese de compostos ativos), e outros elementos genéticos (e.g., regiões regulatórias, fatores de transcrição, *operons*, *transposons*, *enhancers*, plasmídeos, ilhas genômicas e sequências repetitivas como CRISPRs).

Apesar da importância dessa etapa, destaca-se que não existem até então *pipelines* consolidadas na literatura para conduzir todas essas análises. Portanto, a escolha das ferramentas e métodos computacionais para a construção de cada *pipeline* de análise deve ser cuidadosamente planejada e justificada, considerando a natureza dos dados, os requisitos computacionais, como também as vantagens e limitações de cada ferramenta (BHARTI; GRIMM, 2021; LIU et al., 2021).

2.2 REVISITANDO ETAPAS E FERRAMENTAS DE BIOINFORMÁTICA

A partir dos dados gerados como saída da plataforma de sequenciamento, disponibilizados comumente em formato `.fastq.gz`, diferentes etapas de tratamento e manipulação das sequências são realizadas visando extrair informações relevantes que respondam às questões delineadas em um dado estudo. Essas etapas demandam o uso de diversas ferramentas de bioinformática, as quais desempenham um papel crucial para possibilitar a análise e interpretação dos resultados visto o grande volume e a inerente complexidade associada aos dados biológicos. Todavia, decorrente do avanço contínuo das tecnologias de sequenciamento assim como da expansão da capacidade computacional, verifica-se um crescimento significativo no desenvolvimento de métodos e ferramentas computacionais especializadas para cada uma das etapas de operação. Por consequência, a escolha da ferramenta mais adequada exige um planejamento cuidadoso, levando em consideração a natureza dos dados, os requisitos computacionais, assim como a justificativa técnica para a adoção de uma dada ferramenta (com base em suas vantagens e limitações). Ademais, é importante lembrar que a manipulação das amostras nas etapas anteriores impacta diretamente os resultados obtidos nas análises subsequentes, ressaltando assim a importância de uma documentação rigorosa para garantir a confiabilidade e a reproduzibilidade dos resultados obtidos (QUINCE et al., 2017; BHARTI; GRIMM, 2021). Nesse contexto, algumas etapas e ferramentas relevantes são descritas a seguir, as quais são então arranjadas na construção de *pipelines* de acordo com o tipo de análise⁷ a ser realizada.

⁷ Vale destacar que algumas análises são realizadas diretamente sobre as leituras brutas ao passo que outras exigem a montagem prévia das sequências (QUINCE et al., 2017).

2.2.1 Limpeza e controle de qualidade

Independentemente da análise a ser conduzida por uma dada *pipeline*, a implementação de uma etapa de pré-processamento dos dados (brutos) obtidos a partir do sequenciamento da amostra tem se mostrado indispensável para assegurar a qualidade e representatividade das leituras (BHARTI; GRIMM, 2021). A implementação adequada dessa etapa implica, geralmente, em uma redução significativa no número de leituras, o que portanto traz maior eficiência para as etapas subsequentes (LU et al., 2022). Essa etapa inclui, inicialmente, a filtragem de i) adaptadores presentes nas leituras, ii) leituras redundantes, iii) leituras que contêm apenas adaptadores e/ou iv) leituras que possuem baixa qualidade (com muitas bases desconhecidas), o que pode ser realizado através de ferramentas bem estabelecidas como o *CutAdapt* (MARTIN, 2011), *Sickle* (JOSHI; FASS, 2011), *Trimmomatic* (BOLGER; LOHSE; USADEL, 2014) e *AdapterRemoval* (SCHUBERT; LINDGREEN; ORLANDO, 2016). Em seguida, dependendo da origem dos dados (local de coleta da amostra), torna-se pertinente proceder a remoção de possíveis contaminações encontradas nas leituras brutas que sejam provenientes do DNA do hospedeiro, e.g., usando ferramentas como o *Bowtie 2*⁸ (LANGMEAD; SALZBERG, 2012), *DeconSeq* (SCHMIEDER; EDWARDS, 2011) ou *BWA-MEM* (LI, 2013; VASIMUDDIN et al., 2019). Por fim, a determinação de estatísticas quanto à qualidade das leituras deve ser realizada utilizando ferramentas como o *FastQC* (ANDREWS, 2010), *NGS QC Toolkit* (PATTEL; JAIN, 2012), *FQC Dashboard* (BROWN; PIRRUNG; MCCUE, 2017) ou *MultiQC*⁹ (EWELS et al., 2016), para garantir que os dados (limpos) estejam íntegros e adequados para as análises subsequentes. Visando unificar as operações relacionadas à filtragem e ao controle de qualidade, a ferramenta *open-source Fastp* (CHEN, 2023) tem emergido recentemente como uma solução interessante do ponto de vista prático (especialmente, em função do desempenho otimizado, excelente robustez, simplicidade e riqueza de funções), a qual portanto passa a ser adotada aqui.

2.2.2 Classificação taxonômica

A classificação taxonômica visa caracterizar a composição do microbioma de um dado ambiente a partir da observação da abundância e da diversidade das comunidades microbianas presentes na amostra coletada. Essa tarefa, fundamental em muitas *pipelines* de análise de dados de metagenômica, consiste na atribuição de rótulos taxonômicos às sequências (leituras ou *contigs*), i.e., busca-se identificar o táxon associado a cada sequência por comparação/similaridade/alinhamento com bancos de dados de genomas de referência (conhecidos). Para tal, diferentes ferramentas que empregam abordagens distin-

⁸ Basicamente, leituras mapeadas ao genoma de referência do hospedeiro (assumido conhecido) pelo *Bowtie 2* são tratadas como leituras contaminadas e, então, removidas.

⁹ A ferramenta permite agregar resultados de análises de bioinformática de várias amostras em um único relatório, buscando em um diretório específico os *logs* das análises realizadas e compilando em um relatório em HTML.

tas (com suas vantagens e limitações) podem ser utilizadas, tais como **Kraken** (WOOD; SALZBERG, 2014), **CLARK** (OUNIT et al., 2015), **Taxator-tk** (DRÖGE; GREGOR; MCHARDY, 2015), **MetaPhlAn2** (TRUONG et al., 2015), **Centrifuge** (KIM et al., 2016), **Megan6** (HUSON et al., 2007; HUSON et al., 2016), **Taxonomer** (FLYGARE et al., 2016), **Kaiju** (MENZEL; NG; KROGH, 2016), **DUDes** (PONS et al., 2019) ou **Kraken2** (WOOD; LU; LANGMEAD, 2019). Dentre elas, a ferramenta **Kraken2** (WOOD; LU; LANGMEAD, 2019) aliada à **Bracken** (LU et al., 2017) tem se destacado (devido à precisão, eficiência e desempenho) em comparações com outras (YE et al., 2019; SEPPEY; MANNI; ZDOBNOV, 2020; EDWIN et al., 2024). Especificamente, a ferramenta **Kraken2** realiza a classificação taxonômica¹⁰ utilizando um algoritmo que associa subsequências (*substrings*) genômicas curtas (*k-mers*) ao menor ancestral comum (LCA) dos táxons, usando uma tabela de *hash* probabilística compacta construída com base nas sequências de um banco de dados de referência (e.g., o NCBI). Como resultado, obtém-se relatórios detalhados contendo o percentual de fragmentos cobertos pelo grupo taxonômico associado a tal táxon, número de fragmentos cobertos pelo grupo taxonômico, número de fragmentos atribuídos diretamente ao táxon, código de hierarquia taxonômica, ID taxonômico com base no NCBI e nome científico indentado. A partir desses resultados gerados pelo **Kraken2**, o **Bracken** possibilita refinar a estimativa de abundância de espécies usando um modelo Bayesiano; em suma, o **Bracken** reatribui as leituras inicialmente classificadas em níveis taxonômicos mais altos (como gênero ou família) para reestimar com maior precisão a abundância de espécies. Outros *scripts* úteis para a manipulação dos resultados gerados pelo **Kraken2** e/ou **Bracken** são disponibilizados através do pacote **KrakenTools** (KRAKENTOOLS DEVELOPERS, 2021; LU et al., 2022), trazendo funções para gerar gráficos, filtrar e combinar relatórios, bem como calcular métricas de α - e β -diversidade para cada amostra. Por fim, os resultados obtidos da classificação taxonômica das sequências presentes em uma dada amostra podem ser visualizados usando ferramentas como **Krona** (OND OV; BERGMAN; PHILLIPPY, 2011) e **Pavian** (BREITWIESER; SALZBERG, 2020).

2.2.3 Montagem metagenômica

A montagem metagenômica consiste em agrupar/combinar as leituras (pré-processadas e limpas) para gerar fragmentos maiores, denominados *contigs*, visando reconstruir (ao menos parcialmente) genomas dos diferentes microrganismos presentes na amostra. Dentre as diferentes ferramentas disponíveis para essa tarefa, destacam-se a **RayMeta** (BOIS-VERT et al., 2012), **MetaVelvet** (NAMIKI et al., 2012), **SOAPdenovo2** (LUO et al., 2012), **IDBA-UD** (PENG et al., 2012), **Omega** (HAIDER et al., 2014), **MEGAHIT** (LI et al., 2015)

¹⁰ Vale mencionar que um protocolo bastante detalhado é descrito por Lu et al. (2022), o qual faz uso das ferramentas **Bowtie 2** (LANGMEAD; SALZBERG, 2012) para remoção de contaminações, **Kraken2** (WOOD; LU; LANGMEAD, 2019) para classificação taxonômica, **Bracken** (LU et al., 2017) para refinar a abundância em nível de espécies, enquanto **Krona** (OND OV; BERGMAN; PHILLIPPY, 2011) e **Pavian** (BREITWIESER; SALZBERG, 2020) são empregadas para visualização dos resultados.

e **metaSPAdes**¹¹ (NURK et al., 2017). Tais ferramentas, geralmente baseadas em grafos de *Bruijn* (COMPEAU; PEVZNER; TESLER, 2011), fragmentam as leituras em subsequências de tamanho fixo¹² (*k-mers*) que se sobrepõem parcialmente, formando assim os vértices e arestas de um grafo; então, cabe à ferramenta de montagem encontrar um caminho através desse grafo que reconstrua os genomas subjacentes (QUINCE et al., 2017). Todavia, especialmente em amostras complexas (contendo múltiplas comunidades de microrganismos), a operação de montagem enfrenta desafios intrínsecos decorrentes de erros de sequenciamento, da presença de regiões repetitivas (tanto intergenômicas quanto intragenômicas), da coexistência de diferentes linhagens (com genomas semelhantes, mas não idênticos) e da cobertura desigual ao longo de diferentes genomas (NAMIKI et al., 2012; HOWE et al., 2014; ABRAM, 2015). Ademais, a variabilidade de abundância entre as espécies presentes em uma dada amostra (que podem variar de altamente prevalentes a extremamente raras) impõe desafios extras¹³, já que a cobertura de cada microrganismo afeta a qualidade e a extensão dos *contigs* obtidos (AYLING; CLARK; LEGGETT, 2019). Diante disso, torna-se evidente que a escolha adequada da ferramenta de montagem revela-se crucial para que as análises subsequentes sejam mais confiáveis, devendo essa escolha levar em conta o enfoque do estudo, a composição estimada da comunidade microbiana, o grau de complexidade (número de espécies e distribuição de abundâncias) e a disponibilidade de recursos computacionais (VOLLMERS; WIEGAND; KASTER, 2017; WALT et al., 2017; AYLING; CLARK; LEGGETT, 2019). Na prática, apesar de nenhuma ferramenta ter se provado universalmente superior para todos os tipos de dados, as ferramentas **metaSPAdes** e **MEGAHIT** vêm sendo preferidas para lidar com amostras complexas contendo múltiplos microrganismos, já que exibem desempenho superior frente à dados sintéticos e reais (WALT et al., 2017; QUINCE et al., 2017; BHARTI; GRIMM, 2021; LIU et al., 2021; ZHANG et al., 2022). O **metaSPAdes** integra informações de cobertura durante a construção do grafo (NURK et al., 2017), favorecendo assim a obtenção de *contigs* mais longos; contudo, essa abordagem demanda maior quantidade de memória e tempo de processamento. Por sua vez, o **MEGAHIT** utiliza estruturas de dados sucintas para reduzir o consumo de memória, sendo altamente eficiente na montagem a partir de grandes volumes de leituras curtas (LI et al., 2015); apesar de produzir *contigs* ligeiramente mais curtos, essa abordagem destaca-se pela execução mais rápida e pelo menor consumo de recursos. Vale destacar que, especialmente para amostras complexas (e.g., solo e oceano), o **MEGAHIT** demonstrou capacidade superior de montar mais genes que podem ser anotados,

¹¹ Para mais detalhes quanto ao uso da ferramenta, veja os protocolos apresentados recentemente em (PRJIBELSKI et al., 2020).

¹² Tanto o **MEGAHIT** (LI et al., 2015) quanto o **metaSPAdes** (NURK et al., 2017) analisam iterativamente o comprimento do *k-mers* a fim de encontrar o valor ótimo (BHARTI; GRIMM, 2021).

¹³ Em contraste com a montagem de um único genoma, onde se supõe que a cobertura de sequências ao longo do genoma seja aproximadamente uniforme, a montagem metagenômica envolve múltiplos microrganismos, resultando intrinsecamente em coberturas variáveis devido a diferenças na abundância, tamanho dos genomas, conteúdo e vieses de sequenciamento (QUINCE et al., 2017).

o que suporta a sua escolha (QUINCE et al., 2017; ZHANG et al., 2022).

2.2.4 Predição de genes

A predição de genes a partir de fragmentos de DNA, oriundos das sequências limpas ou *contigs* gerados na etapa de montagem, representa uma etapa fundamental para elucidar o potencial funcional das comunidades microbianas presentes em uma dada amostra. Diversos métodos têm sido amplamente empregados para identificar marcos de leitura (*open reading frames*, ORFs) codificadores de proteínas, sendo eles geralmente categorizados em i) abordagens baseadas em homologia, ii) métodos fundamentados em modelos (e.g., modelos de *Markov*) e iii) técnicas de *machine learning* (ZHANG; JIN; ZHANG, 2017). Ferramentas bem estabelecidas, tais como o **GeneMarkS** (BESEMER; LOMSADZE; BORODOVSKY, 2001), o **Glimmer3** (DELCHER et al., 2007) e o **Prodigal** (HYATT et al., 2010), exibem acurácia acima de 97% na detecção de ORFs verdadeiramente codificantes (HYATT et al., 2010; DELCHER et al., 2007; BESEMER; LOMSADZE; BORODOVSKY, 2001). Contudo, a identificação dos sítios de início das ORFs ainda apresenta certa imprecisão, devido (em parte) à dificuldade em detectar genes cujos padrões de sequência não correspondem adequadamente a um modelo específico de espécie (HYATT et al., 2010; BORODOVSKY et al., 1995). Já no contexto de metagenômica, genes presentes nos *contigs* resultantes da montagem podem ser anotados usando ferramentas como o **metaGeneMark**¹⁴ (ZHU, 2010) e o **Prokka** (SEEMANN, 2014). Dentre elas, destaca-se a ferramenta **Prokka** (adotada aqui), a qual foi desenvolvida especificamente para ser precisa e rápida, já que permite processar paralelamente várias subseções dos dados. Essa ferramenta incorpora uma série de outras para realizar a anotação gênica (e.g., o **Prodigal**), fazendo com que operações diversificadas sejam executadas de maneira simples e direta (SEEMANN, 2014). Cabe salientar que, na prática, a escolha da ferramenta deve levar em conta a variedade e complexidade das espécies presentes na amostra, a disponibilidade de modelos genômicos de referência e as restrições computacionais, de modo a maximizar a qualidade das anotações funcionais que embasarão as análises subsequentes.

2.2.5 Identificação de genes

Esta etapa tem como objetivo identificar genes presentes em sequências nucleotídicas ou proteicas, tendo como referência um banco de dados de genes conhecidos. Para realizar essa operação, que comumente envolve o inerente uso de ferramentas de alinhamento [e.g., o **BLAST** (ALTSCHUL et al., 1990) e o **KMA** (CLAUSEN; AARESTRUP; LUND, 2018)], diferentes *softwares* podem ser empregados. Dentre eles, se destacam o **ABRicate**

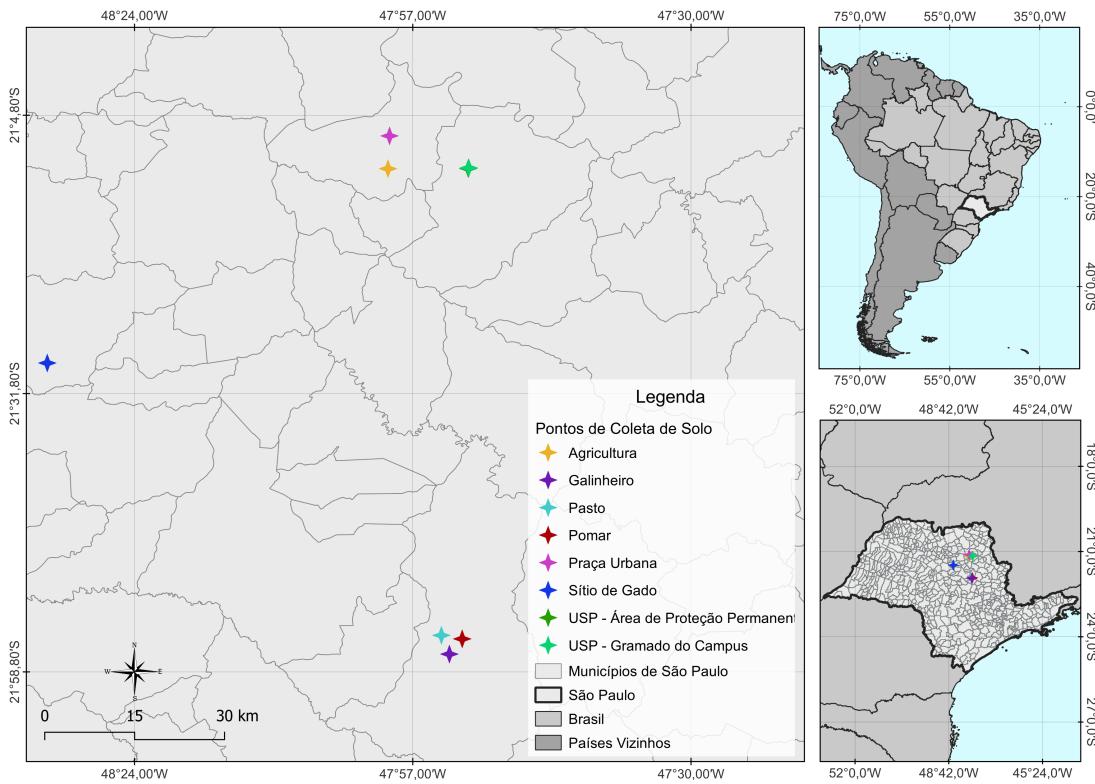
¹⁴ Recentemente, a ferramenta **MetaGeneMark-2** tem sido desenvolvida para aprimorar, de forma significativa, a acurácia na identificação de genes em *contigs* de genomas complexos, demonstrando melhor desempenho em comparação com outras ferramentas de última geração; sobretudo, na predição de sítios de início (GEMAYEL; LOMSADZE; BORODOVSKY, 2022).

(SEEMANN, 2020), o AMRFinderPlus (FELDGARDEN et al., 2021), o ResFinder (FLORENSA et al., 2022), o Resistance Gene Identifier (RGI) (ALCOCK et al., 2023) e o ARGs-OAP (YIN et al., 2023). Todavia, visto que a execução desses softwares depende da disponibilidade de bancos de dados de referência contendo sequências gênicas conhecidas (já anotadas), a escolha do banco de dados mais adequado torna-se crucial e deve ser realizada de acordo com a categoria do perfil gênico de interesse. Particularmente com respeito à identificação de ARGs e de VFs, verifica-se que os seguintes bancos de dados vêm sendo amplamente utilizados: Virulence Factor Database (VFDB) (CHEN et al., 2005), MvirDB (ZHOU et al., 2007), Antibiotic Resistance Genes Database (ARDB) (LIU; POP, 2009), Comprehensive Antibiotic Resistance Database (CARD) (MCARTHUR et al., 2013), Victors (SAYERS et al., 2019), ResFinder database (BORTOLAIA et al., 2020) e structured ARG (SARG) database (YIN et al., 2023). Diante do exposto, considera-se aqui o uso do software ABRicate, desenvolvido especificamente para a identificação do resistoma e do viruloma presentes em arquivos de sequências nucleotídicas (SEEMANN, 2020), associado aos bancos de dados de referência CARD (MCARTHUR et al., 2013) e VFDB (CHEN et al., 2005). Vale reforçar que esses bancos de dados são uma escolha confiável, visto que são constantemente curados e atualizados, além de estarem prontamente disponíveis (de forma nativa) no software considerado.

2.3 DESCRIÇÃO DO CONJUNTO DE DADOS

De acordo com Ordine et al. (2023), as amostras de solo foram obtidas de 8 diferentes pontos/sítios de coleta localizados na região nordeste do Estado de São Paulo, nas proximidades das cidades de Ribeirão Preto, Sertãozinho e São Carlos (conforme ilustrado na Figura 1). Cada um dos pontos/sítios de coleta apresenta características significativamente distintas quanto à classificação geológica, atividade antrópica e localização (como sintetizado na Tabela 1). Especificamente, em cada um dos 8 pontos/sítios de coleta, 3 subamostras de solo foram obtidas, consistindo da retirada de aproximadamente 50 g de solo dos primeiros 10 cm a partir da superfície após a remoção de 5 cm de serrapilheira (folhas, galhos e outros detritos vegetais). Essas subamostras foram misturadas para obter uma melhor representação das comunidades microbianas do local. Parte dessa mistura foi armazenada em tubos Falcon estéreis para posterior extração do DNA. O DNA de cada amostra foi extraído usando o kit *DNeasy PowerSoil* da QIAGEN, seguindo às recomendações do fabricante. A quantificação e a verificação de qualidade do material foram realizadas com o *Nanodrop One* da Thermo Fisher Scientific, em conjunto com um gel de agarose (1%) exposto à eletroforese. O material foi, então, submetido ao sequenciamento metagenômico (do tipo *shotgun*) utilizando a plataforma Illumina NovaSeq 6000, considerando uma profundidade de 12 Gb por amostra. Os dados brutos (i.e., *raw sequence reads* em formato *.fastq*), obtidos do sequenciamento, se fazem disponíveis no

Figura 1 – Localização geográfica, no estado de São Paulo, dos pontos/sítios de coleta das amostras de solo consideradas [conforme (ORDINE et al., 2023, Table 1)].



Fonte: Autoria própria.

repositório *Sequence Read Archive* (SRA), mantido pelo *National Center for Biotechnology Information* (NCBI), sob o número do *BioProject PRJNA900430* (ORDINE et al., 2023). Vale mencionar que os metadados disponíveis nas amostras são a data da coleta da amostra, bem como a latitude e a longitude dos pontos/sítios de coleta.

Dante do exposto até aqui, pode-se agora proceder para a construção e implementação das *pipelines* necessárias para conduzir a análise dos dados de metagenômica pertinentes ao presente trabalho.

Tabela 1 – Principais informações relacionadas às amostras de solo consideradas.

Código e denominação da amostra	Classificação geológica	Síntese da atividade antrópica	Sítio e data de coleta
SRR22278225 (Floresta)	Latossolo vermelho	Área de proteção permanente com acesso restrito à pesquisas.	-21.1662, -47.86036 (01/03/2020)
SRR22278226 (Agricultura)	Latossolo vermelho	Grande propriedade agrícola com cultivo de cana-de-açúcar e uso de agroquímicos.	-21.16643, -47.99004 (25/11/2019)
SRR22278227 (Pomar)	Neossolo quartzarênico	Pequena propriedade de agricultura familiar com pomar.	-21.92674, -47.87008 (01/09/2020)
SRR22278228 (Pastagem)	Neossolo quartzarênico	Pequena propriedade familiar de agropecuária (criação de gado).	-21.921, -47.90373 (01/09/2020)
SRR22278229 (Praça)	Latossolo vermelho	Área urbana com alta circulação de pessoas e pequenos animais.	-21.1134, -47.98762 (01/09/2020)
SRR22278230 (Gramado)	Latossolo vermelho	Área urbana com alta circulação de pessoas e pequenos animais.	-21.16511, -47.85944 (01/03/2020)
SRR22278231 (Criação de Gado)	Arenito Bauru	Grande propriedade agropecuária com alta circulação de pessoas e animais.	-21.48066, -48.54118 (01/04/2020)
SRR22278232 (Aviário)	Neossolo quartzarênico	Pequena propriedade de agricultura familiar com criação de aves.	-21.95133, -47.89079 (01/09/2020)

Fonte: Adaptado de (ORDINE et al., 2023, Table 1).

3 PIPELINES PARA PROCESSAMENTO E ANÁLISE DE DADOS

Nesta seção, as *pipelines* desenvolvidas para conduzir a análise de dados de metagenômica [disponibilizados por Ordine et al. (2023)] são apresentadas (conforme ilustrado na Figura 2). Especificamente, uma das *pipelines* visa a classificação taxonômica das comunidades microbianas presentes nas diferentes amostras [Figura 2(a)], a qual é baseada no protocolo descrito em (LU et al., 2022) compreendendo as seguintes etapas: 1) o controle de qualidade e limpeza dos dados brutos, usando o *software Fastp* (CHEN, 2023); 2) a classificação taxonômica propriamente dita a partir de sequências filtradas, realizada pelo *software Kraken2* (WOOD; LU; LANGMEAD, 2019); e 3) o refinamento dos resultados obtidos, através do *software Bracken* (LU et al., 2017). Já a outra *pipeline*, baseada na descrição apresentada em Ordine et al. (2023), objetiva a caracterização do resistoma e do viruloma através da identificação de ARGs e VFs [Figura 2(b)], sendo composta pelas seguintes etapas: 1) o controle de qualidade e limpeza dos dados brutos, usando o *software Fastp* (CHEN, 2023); 2) montagem de *contigs* a partir das sequências filtradas, usando o *software MEGAHIT* (LI et al., 2015); 3) predição de genes, utilizando o *software Prokka* (SEEMANN, 2014); e, por fim, 4) a identificação de genes a partir do *software ABRicate* (SEEMANN, 2020), associado aos bancos de dados CARD (MCARTHUR et al., 2013) e VFDB (CHEN et al., 2005). Diante disso, são apresentados a seguir os *scripts*¹ desenvolvidos para a obtenção e processamento dos dados de metagenômica, em linguagem GNU Bash (GNU, 2007), bem como os códigos implementados para realizar as análises dos resultados, em linguagem Python (ROSSUM; DRAKE, 2009).

3.1 OBTENÇÃO DOS DADOS DE METAGENÔMICA (AMOSTRAS)

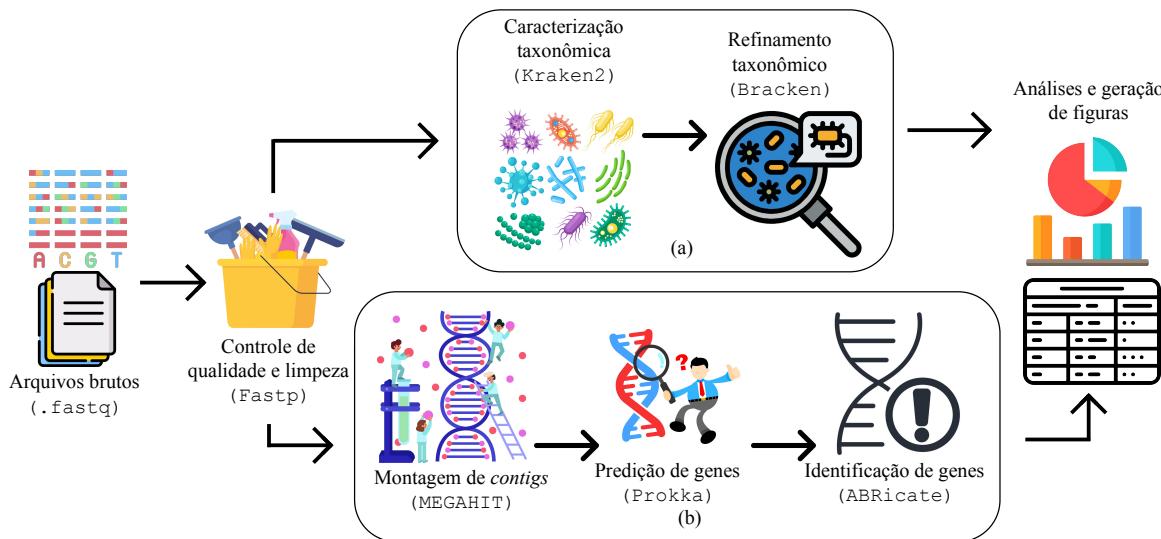
O *Script* 1 implementa a etapa de obtenção dos dados brutos de metagenômica (em formato `.fastq`) armazenados no SRA, empregando para isso o *software SRA toolkit* (NCBI, 2024b). Esse *script* pode ser dividido em três fragmentos, a saber:

- **Linhas 1–12:** definição de algumas variáveis e hiperparâmetros importantes, incluindo o caminho do diretório de armazenamento dos dados;
- **Linhas 14–21:** função responsável por efetivamente obter os dados do SRA, fazendo uso do *SRA toolkit*;
- **Linhas 23–27:** laço de repetição que opera sobre cada valor de amostra definida em `accessions`, evocando a função `run_sra_toolkit()`.

Ao final da execução desse *script*, são gerados 16 arquivos com extensão `.fastq` (i.e., 2 arquivos por amostras) no diretório definido em `fastq_dir`.

¹ Para detalhes, veja <https://github.com/lablapse/metagenomics>.

Figura 2 – Diagrama de blocos ilustrando as *pipelines* desenvolvidas. (a) Etapas exclusivas à *pipeline* de classificação taxonômica. (b) Etapas exclusivas à *pipeline* de caracterização do resistoma e do viruloma.



Fonte: Autoria própria.

3.2 CONTROLE DE QUALIDADE E LIMPEZA

O *Script 2* implementa a operação de controle de qualidade e limpeza sobre os dados brutos obtidos do SRA, utilizando o *software* *Fastp* (discutido na Seção 2.2.1). Nesse *script*, destacam-se os seguintes pontos:

- **Linhas 1–9:** definição de algumas variáveis importantes, tais como o local de armazenamento dos arquivos brutos e o diretório de destino dos resultados;
- **Linhas 11–39:** função que encapsula as operações de controle de qualidade e limpeza, definindo variáveis locais usadas pelo *Fastp*;
- **Linhas 41–46:** evoca a função `run_fastp()` para cada amostra definida em `accessions`.

Ao final da execução desse *script*, são gerados 32 arquivos no diretório definido em `output_dir`, sendo 16 no formato `.fastq.gz` (dados filtrados e compactados), 8 em formato `.html` e 8 em formato `.json`.

Script 1 – Obtenção dos dados de metagenômica (amostras) do SRA.

```

1 #!/bin/bash
2 #SRA toolkit: v3.1.1
3 prefetch_dir="/sra/prefetch/dir/"
4 export NCBI_SRA_DIR=$prefetch_dir
5
6 scratch_path="/scratch/dir/"
7 fastq_dir="/saving/fastq/dir/"
8
9 accessions=(SRR22278225 SRR22278226 SRR22278227 SRR22278228 \
10           SRR22278229 SRR22278230 SRR22278231 SRR22278232)
11
12 num_threads=$(nproc)
13
14 run_sra_toolkit() {
15   local accession=$1
16   prefetch ${accession}
17   fasterq-dump -t ${scratch_path} \
18     -e ${num_threads} \
19     -O ${fastq_dir} \
20     ${accession}
21 }
22
23 for accession in "${accessions[@]}"
24 do
25   echo "Processando ${accession}..."
26   run_sra_toolkit "$accession"
27 done

```

Fonte: Autoria própria.

3.3 CLASSIFICAÇÃO TAXONÔMICA

O *Script 3* implementa a operação de classificação taxonômica das sequências filtradas, utilizando o *software Kraken2* (discutido na Seção 2.2.2). Nesse *script*, cabe destacar as seguintes partes:

- **Linhas 1–11:** definição de algumas variáveis importantes, incluindo o diretório com o banco de dados a ser utilizado;
- **Linhas 13–39:** laço de repetição que opera sobre cada valor de amostra definida em `accessions`, definindo variáveis locais utilizadas durante a execução do *Kraken2*.

Ao final da execução desse *script*, são gerados, dentro do diretório definido em `output_dir`, 8 arquivos de extensão `.tsv` detalhando a abundância dos diferentes microorganismos identificados na amostra, 8 arquivos de extensão `.kraken` contendo detalhes das sequências identificadas, 16 arquivos de extensão `.fastq.gz` listando as sequências classificadas e, ainda, 16 arquivos de extensão `.fastq.gz` listando as sequências não classificadas.

Script 2 – Controle de qualidade e limpeza dos dados de metagenômica.

```

1 #!/bin/bash
2 #Fastp: v0.23.4
3 fastq_dir="/stored/fastq/dir/"
4 output_dir="/saving/filtered_fastq/dir/"
5
6 accessions=(SRR22278225 SRR22278226 SRR22278227 SRR22278228 \
7             SRR22278229 SRR22278230 SRR22278231 SRR22278232)
8
9 num_threads=$(nproc)
10
11 run_fastp(){
12     local accession=$1
13     local shared_input="${fastq_dir%}/ ${accession}"
14     local shared_output="${output_dir%}/ ${accession}"
15
16     local forward_read="${shared_input}_1.fastq"
17     local reverse_read="${shared_input}_2.fastq"
18     local forward_filtered="${shared_output}_1_filtered.fastq.gz"
19     local reverse_filtered="${shared_output}_2_filtered.fastq.gz"
20     local html_report="${shared_output}.html"
21     local json_report="${shared_output}.json"
22
23     echo "Executando o Fastp para ${accession}..."
24     fastp \
25         -i "${forward_read}" \
26         -I "${reverse_read}" \
27         -o "${forward_filtered}" \
28         -O "${reverse_filtered}" \
29         -h "${html_report}" \
30         -j "${json_report}" \
31         --thread ${num_threads} \
32         --correction \
33         --trim_poly_g \
34         --trim_poly_x \
35         --qualified_quality_phred 30 \
36         --length_required 50 \
37         --low_complexity_filter \
38         --dedup
39 }
40
41 for accession in "${accessions[@]}"
42 do
43     echo "Processando ${accession}..."
44     run_fastp "${accession}"
45     echo "${accession}: processamento completo!"
46 done

```

Fonte: Autoria própria.

Script 3 – Classificação taxonômica sobre os dados filtrados.

```

1 #!/bin/bash
2 # Kraken2: v.2.1.2
3
4 input_dir="/stored/fastq/filtered/dir"
5 output_dir="/saving/taxonomy/dir"
6 custom_database="/custom/database/dir"
7
8 accessions=(SRR22278225 SRR22278226 SRR22278227 SRR22278228 \
9           SRR22278229 SRR22278230 SRR22278231 SRR22278232)
10
11 num_threads=$(nproc)
12
13 for accession in "${accessions[@]}"
14 do
15   echo "Processando ${accession}"
16
17   input_r1="${input_dir}/${accession}_1_filtered.fastq.gz"
18   input_r2="${input_dir}/${accession}_2_filtered.fastq.gz"
19
20   report_file="${output_dir}/${accession}_report.tsv"
21   kraken_file="${output_dir}/${accession}_report.kraken"
22   classified_file="${output_dir}/${accession}_classified_.fastq.gz"
23   unclassified_file="${output_dir}/${accession}_unclassified_.fastq.gz"
24
25   echo "Executando Kraken2 para ${accession}"
26
27   kraken2 --memory-mapping \
28     --threads "${num_threads}" \
29     --use-names \
30     --gzip-compressed \
31     --db "${custom_database}" \
32     --paired "${input_r1}" "${input_r2}" \
33     --report "${report_file}" \
34     --output "${kraken_file}" \
35     --classified-out "${classified_file}" \
36     --unclassified-out "${unclassified_file}"
37
38   echo "${accession}: Processamento completo!"
39 done

```

Fonte: Autoria própria.

3.4 MONTAGEM DE CONTIGS

O *Script 4* implementa a operação de montagem de *contigs* a partir dos dados provenientes da etapa de controle de qualidade e limpeza, utilizando o *software* MEGAHIT (discutido na Seção 2.2.3). Nesse *script*, os seguintes pontos podem ser destacados:

- **Linhas 1–12:** definição de variáveis importantes e criação do diretório para o armazenamento dos diversos arquivos gerados durante a montagem;
- **Linhas 14–28:** função que encapsula a operação de montagem dos dados, definindo variáveis locais utilizadas durante a execução do MEGAHIT;

- Linhas 30–34: evoca a função `run_megahit()` para cada amostra definida em `accessions`.

Ao final da execução desse *script*, são gerados 8 diretórios em `output_megahit_dir`, i.e., um para cada amostra; em cada diretório, tem-se um arquivo `final.contigs.fa` contendo os *contigs* montados.

Script 4 – Montagem dos *contigs* pelo software MEGAHIT.

```

1 #!/bin/bash
2 #MEGAHIT: v1.2.9
3
4 filtered_dir="/stored/fastq/filtered/dir/"
5 output_megahit_dir="/saving/assembled/dir/"
6
7 accessions=(SRR22278225 SRR22278226 SRR22278227 SRR22278228 \
8           SRR22278229 SRR22278230 SRR22278231 SRR22278232)
9
10 num_threads=$(nproc)
11
12 mkdir -p "${output_megahit_dir}"
13
14 run_megahit() {
15   local accession=$1
16   local shared_input="${filtered_dir%}/$accession"
17
18   local forward_filtered="${shared_input}_1_filtered.fastq.gz"
19   local reverse_filtered="${shared_input}_2_filtered.fastq.gz"
20   local output_dir="${output_megahit_dir%}/$accession"
21
22   echo "Executando MEGAHit para amostra ${accession}..."
23   megahit \
24     -1 "${forward_filtered}" \
25     -2 "${reverse_filtered}" \
26     -o "${output_dir}" \
27     --num-cpu-threads "${num_threads}"
28 }
29
30 for accession in "${accessions[@]}"; do
31   echo "Processando ${accession}..."
32   run_megahit "$accession"
33   echo "${accession}: processamento completo."
34 done

```

Fonte: Autoria própria.

3.5 PREDIÇÃO DE GENES

O *Script 5* implementa a operação de predição de genes sobre os *contigs* montados obtidos na saída da etapa anterior, utilizando para isso o *software Prokka* (discutido na Seção 2.2.4). Nesse *script*, destacam-se as seguintes partes:

- Linhas 1–12: definição de algumas variáveis e criação do diretório para o armazenamento dos arquivos gerados pelo processo de anotação gênica;

- **Linhas 14–25:** função que encapsula a operação de predição de genes, considerando alguns parâmetros importantes utilizadas pelo Prokka dada a natureza dos dados;
- **Linhas 27–31:** evoca a função `run_prokka()` para cada amostra definida em `accessions`.

Ao final da execução desse *script*, são gerados 8 diretórios em `output_prokka_dir`, um para cada amostra; em cada diretório, o Prokka adiciona vários arquivos, sendo os dados de genes preditos armazenados em um arquivo de anotação com extensão `.gff`.

Script 5 – Predição de genes a partir dos *contigs* montados.

```

1 #!/bin/bash
2 #Versao do Prokka: 1.14.6
3
4 assembled_dir="/stored/assembled/contigs/dir/"
5 output_prokka_dir="/saving/prediction/dir/"
6
7 accessions=(SRR22278225 SRR22278226 SRR22278227 SRR22278228 \
8           SRR22278229 SRR22278230 SRR22278231 SRR22278232)
9
10 num_threads=$(nproc)
11
12 mkdir -p "${output_prokka_dir}"
13
14 run_prokka() {
15   local accession=$1
16   ln -fs "${assembled_dir%}/{$accession}/final.contigs.fa"
17   local output_dir="${output_prokka_dir%}/{$accession}"
18
19   echo "Executando Prokka ${accession}..."
20   prokka final.contigs.fa \
21     --kingdom Bacteria \
22     --outdir "${output_dir}" \
23     --cpus "${num_threads}" \
24     --metagenome
25 }
26
27 for accession in "${accessions[@]}"; do
28   echo "Processando ${accession}..."
29   run_prokka "${accession}"
30   echo "${accession}: processamento completo."
31 done

```

Fonte: Autoria própria.

3.6 IDENTIFICAÇÃO DE GENES DE INTERESSE

O *Script* 6 define a operação de identificação dos genes de interesse tendo como entrada os genes preditos pelo Prokka, onde se utiliza o *software* ABRicate (discutido na Seção 2.2.5) em conjunto com os bancos de dados CARD e VFDB (ambos atualizados em 29/01/2025). Nesse *script*, cabe destacar os seguintes aspectos:

- **Linhas 1–13:** definição de algumas variáveis relevantes e dos bancos de dados de referência, bem como criação do diretório para o armazenamento dos arquivos gerados pelo processo de identificação de genes;
- **Linhas 15–30:** função que encapsula a operação de identificação de genes de interesse, definindo as variáveis de operação utilizadas pelo ABRicate;
- **Linhas 32–36:** evoca a função `run_abricate()` para cada amostra definida em `accessions`;

Ao final da execução desse *script*, são gerados 8 diretórios em `output_abricate_dir` (um por amostra analisada); em cada diretório, tem-se 2 arquivos (nomeados como `card.tsv` e `vfdb.tsv`) contendo os ARGs e VFs identificados.

Script 6 – Identificação de genes de interesse a partir de dados anotados.

```

1 #!/bin/bash
2 #ABRicate: v1.0.1
3
4 predicted_dir="/stored/predicting/dir/"
5 output_abricate_dir="/saving/identification/dir/"
6
7 num_threads=8
8
9 accessions=(SRR22278225 SRR22278226 SRR22278227 SRR22278228 \
10           SRR22278229 SRR22278230 SRR22278231 SRR22278232)
11 databases=("card" "vfdb")
12
13 mkdir -p "${output_abricate_dir}"
14
15 run_abricate(){
16   local accession=$1
17   local prokka_file=("${predicted_dir%}/$accession/*.gff")
18   local output_dir="${output_abricate_dir%}/$accession"
19   mkdir -p "$output_dir"
20
21   for database in "${databases[@]}"; do
22     echo "Processando ABRicate para ${database}..."
23     abricate --db "${database}" \
24               --threads "${num_threads}" \
25               --minid 80 \
26               --mincov 80 \
27               "${prokka_file[0]}" > "${output_dir}/${database}.tsv";
28     echo "${database}: Processamento completo."
29   done
30 }
31
32 for accession in "${accessions[@]}"; do
33   echo "Executando ABRicate para $accession..."
34   run_abricate "$accession"
35   echo "$accession: processamento completo."
36 done

```

Fonte: Autoria própria.

3.7 MANIPULAÇÃO DOS ARQUIVOS GERADOS PELO KRAKEN2

Os *Scripts* 7 a 9 são empregados para organizar os arquivos `.tsv` gerados pelo `Kraken2` durante a *pipeline* de classificação taxonômica, facilitando a posterior manipulação. Especificamente, o *Script* 7 é responsável por gerar um arquivo em formato `.csv` contendo todos os níveis taxonômicos presentes no arquivo `inspect_file`. Então, tal arquivo é utilizado pelo *Script* 8 como uma base para a criação de um arquivo mais completo, contendo também as abundâncias referentes aos táxons classificados nas diferentes amostras. Por fim, o *Script* 9, que evoca os anteriores, estrutura o arquivo que contém todas as informações taxonômicas encontradas, as quais servem de base para a construção dos gráficos. Ainda, os *Scripts* 10 e 11 são responsáveis por gerar as figuras que ilustram os principais gêneros taxonômicos caracterizados (veja à frente a Figura 4). Em particular, o *Script* 10 define, entre as linhas 16 e 26, o nível taxonômico analisado, a variável indicando a quantidade máxima dos principais gêneros a serem observados, formata os dados para um padrão específico, insere uma coluna contendo os valores médios das amostras e evoca a função responsável por construir as figuras. Por sua vez, o *Script* 11 implementa a função de geração dos gráficos de barras com a linha sombreada referente aos valores médios. Nessa função, as colunas de interesse são selecionadas e convertidas para valores percentuais, os principais gêneros são selecionados, o gráficos de barras e a linha sombreada referente aos valores médios são estruturados e a figura é salva em formato `.pdf`.

3.8 MANIPULAÇÃO DOS ARQUIVOS GERADOS PELO ABRICATE

Os *Scripts* 12 a 15 ilustram alguns fragmentos de códigos responsáveis por gerar as figuras que mostram os ARGs e os VFs identificados (veja à frente as Figuras 5 e 6). Especificamente, o *Script* 12 realiza duas operações, sendo uma sobre os resultados presentes no arquivo `card.tsv` e outra sobre os resultados do arquivo `vfdb.tsv`. Esse *script* é responsável por evocar as funções de importação e formatação dos dados, de adicionar uma coluna com as cores referentes a cada gene ao *dataframe* completo (para manter a consistência de cores entre o gráfico de barras e a legenda) e de definir o local para o armazenamento das figuras. O *Script* 13 implementa a operação de importação dos arquivos `.tsv`, agrupando os genes identificados, criando uma coluna com o número de instâncias em que os genes foram observados e substituindo as nomenclaturas gênicas longas por sinônimos mais curtos. Ao final, a coluna com as instâncias é convertida para valores percentuais. O *Script* 14 cria um *dataframe* contendo todas as informações gênicas das amostras, provendo assim uma estrutura consistente para facilitar a construção dos gráficos de barras. Por fim, o *Script* 15 implementa a função para a geração das figuras, criando os gráficos de barras e as legendas com as cores apropriadas para cada gene, formatando a figura para o padrão desejado e salvando as figuras no local definido em `path`.

Os outros gráficos de barras apresentados no decorrer do presente trabalho seguiram um padrão de implementação parecido, com a estrutura geral se mantendo semelhante ao apresentado aqui.

Script 7 – Criação do arquivo .csv contendo todos os níveis taxonômicos no banco de dados utilizado pelo Kraken2.

```

1 def parse_kraken_inspect(inspect_file: str, parsed_inspect_file: str):
2     output_line_values = [""] * (len(_BIO_CLASSES) - 1)
3     previous_class_names = [""] * (len(_BIO_CLASSES) - 1)
4     previous_class = None
5     header = ";" .join(_BIO_CLASSES._member_names_[1:])
6
7     with open(inspect_file, "r") as input_file, \
8         open(parsed_inspect_file, "w") as output_file:
9         output_file.write(header + "\n")
10        for line in input_file:
11            line_elements = line.split("\t")
12            actual_class, is_subrank = _check_bio_class(line_elements)
13
14            # Reset on encountering a subroot (e.g., plasmids R1, R2)
15            if actual_class == _BIO_CLASSES.NONE:
16                output_line_values = [""] * (len(_BIO_CLASSES) - 1)
17                previous_class_names = [""] * (len(_BIO_CLASSES) - 1)
18                previous_class = actual_class
19                continue
20
21            # Clear taxonomic levels below the current rank
22            if previous_class != None \
23            and actual_class.value < previous_class.value:
24                output_line_values[actual_class.value:] = [""] * (
25                    len(output_line_values) - actual_class.value
26                )
27
28            # Update taxonomic level values
29            taxon_name = line_elements[5].strip()
30            if not is_subrank:
31                output_line_values[actual_class.value] = taxon_name
32                previous_class_names[actual_class.value] = taxon_name
33                output_file.write(";" .join(output_line_values) + "\n")
34            else:
35                output_line_values[actual_class.value] =
36                    previous_class_names[actual_class.value]
37
38            previous_class = actual_class

```

Fonte: Autoria própria.

Script 8 – Criação do arquivo .csv contendo os níveis taxonômicos e as correspondentes abundâncias de cada amostra.

```

1 def fill_inspect_df(df: pd.DataFrame, report_numbers: int | list,
2     report_files_path: str) -> pd.DataFrame:
3     report_numbers = [report_numbers] \
4         if isinstance(report_numbers, int) \
5         else report_numbers
6
7     df_copy = df.copy(deep=True)
8     df_copy[report_numbers] = 0
9
10    for report_number in report_numbers:
11        report_path = f"{report_files_path}/{report_number}.tsv"
12        with open(report_path, "r") as report_file:
13            for line in report_file:
14                line_elements = line.split("\t")
15                taxon_level, is_subrank = _check_bio_class(
16                    line_elements
17                )
18
19                if taxon_level == _BIO_CLASSES.NONE:
20                    continue
21
22                scientific_name = line_elements[5].strip()
23                slices_list = create_slices(
24                    taxon_level,
25                    scientific_name
26                )
27
28                # Insert the read count into the DataFrame
29                if not is_subrank:
30                    count = int(line_elements[1])
31                    df_copy.loc[tuple(slices_list), report_number] =
32                        count
33
34    return df_copy
35
36 def create_slices(taxon_level, scientific_name):
37     slices = [slice(None)] * (len(_BIO_CLASSES) - 1)
38     slices[taxon_level.value] = scientific_name
39     slice_size = len(slices) - taxon_level.value - 1
40     slices[taxon_level.value + 1:] = [""] * (slice_size)
41
42     return slices

```

Fonte: Autoria própria.

Script 9 – Código responsável por evocar as funções implementadas para gerar o arquivo .csv com as informações de taxonomia das diferentes amostras.

```

1 import pandas as pd # conda install anaconda::pandas
2
3 parse_kraken_inspect(inspect_file="./diretorio/kraken_inspect.tsv",
4                         parsed_inspect_file="./kraken_inspect_parsed.csv")
5 df = pd.read_csv("./kraken_inspect_parsed.csv",
6                   index_col=[i for i in range(0,8)], sep=";",
7                   keep_default_na=False)
8 df.sort_index(inplace=True)
9
10 report_numbers = ["kraken_inspect",
11                     "SRR22278225_report", "SRR22278226_report",
12                     "SRR22278227_report", "SRR22278228_report",
13                     "SRR22278229_report", "SRR22278230_report",
14                     "SRR22278231_report", "SRR22278232_report",]
15 df = fill_inspect_df(df, report_numbers=report_numbers,
16                       report_files_path="./diretorio/reports/")
17
18 report_numbers.pop(0)
19 for report in report_numbers:
20     df = df.rename(columns={report:report[3:11]})
```

df.to_csv("./outputs/complete.csv")

Fonte: Autoria própria.

Script 10 – Trecho de código que evoca as funções responsáveis por gerar os gráficos de barras dos gêneros taxonômicos caracterizados.

```

1 import polars as pl # conda install conda-forge::polars
2
3 def formatting_df(df, taxon):
4     new_df = df.select(
5         pl.col(taxon), pl.col(pl.Int64)
6         ).group_by(pl.col(taxon))
7         .agg(pl.all().max())
8         .drop_nulls()
9
10    new_df = new_df.with_columns(
11        new_df.select(pl.col(pl.Int64).exclude("kraken_inspect"))
12        .mean_horizontal().alias("MEAN")
13        )
14    return new_df
15
16 if __name__ == "__main__":
17     path = "./outputs/complete.csv"
18     df = pl.read_csv(path)
19     taxon = "GENUS"
20     top_value = 15
21     new_df = formatting_df(df, taxon)
22
23     ylim = 15
24     for i in range(25,33):
25         report = f"222782{i}"
26         main_plot(new_df, report, top_value, ylim)
```

Fonte: Autoria própria.

Script 11 – Definição da função responsável por gerar os gráficos de barras dos gêneros taxonômicos caracterizados.

```

1 def main_plot(df, report, top_value, ylim):
2     columns = df.columns
3     taxon = columns[0]
4     mean_sum_values = df["MEAN"].sum()
5     report_sum_values = df[report].sum()
6
7     df = df.top_k(top_value, by="MEAN")
8     df = df.select(
9         pl.col(taxon),
10        pl.col("MEAN") / mean_sum_values * 100,
11        pl.col(report) / report_sum_values * 100,
12    )
13
14     quantity = df.height
15     colors = [(0.5, 0.5, 0.5) for _ in range(quantity)]
16     taxon_names = df[taxon]
17
18     x, y = making_shade_line_to_plot(df["MEAN"])
19     fig, ax = plt.subplots()
20     plt.rcParams["hatch.linewidth"] = 0.4
21     ax.set(xlim=(-1, len(df["MEAN"]) + 0.04))
22     ax.bar(taxon_names, df[report],
23             color=colors, zorder=2, alpha=0.85)
24     ax.plot(
25         x, y,
26         color="black", linewidth=0.4, zorder=1,
27     )
28     ax.fill_between(x, y, color="black", alpha=0.08)
29     ax.set_ylim(0, ylim)
30     ax.set_ylabel("Abundancia (%)")
31     ax.tick_params(axis="x", labelrotation=90)
32     fig = format_figure(fig)
33     fig.savefig(
34         f"plots/{report}.pdf",
35     )
36     return

```

Fonte: Autoria própria.

Script 12 – Trecho de código responsável por evocar as funções de importação dos dados e de geração de figuras acerca dos genes identificados.

```
1 if __name__ == "__main__":
2     databases = ["card", "vfdb"]
3     for database in databases:
4         df_list = []
5         for report in range(22278225, 22278233):
6
7             df2 = load_and_format(f"diretorio/dos/reports/SRR{report}/"\ \
8                                   f"{database}.tsv")
9             df_list.append(df2)
10
11     df = big_dataframe(df_list)
12     genes_guia = df.columns[0]
13     instances_ordine = df.columns[1]
14
15     df = df.sort(
16         instances_ordine, genes_guia,
17         descending=[True, False], nulls_last=True
18     )
19
20     df = df.with_columns(pl.col(pl.Float64) * 100)
21     df = adding_colors_column(df, "#3D9AA6")
22     main_plot(
23         df, f"/diretorio/para/salarvar/as/figuras/"\
24               f"{database}.pdf"
25     )
```

Fonte: Autoria própria.

Script 13 – Definição da função responsável pela importação e formatação dos arquivos .tsv gerados durante a etapa de identificação de genes.

```

1 import polars as pl # conda install conda-forge::polars
2
3 def load_and_format(path):
4     df = pl.read_csv(path, separator="\t")
5     df = df.group_by("GENE").agg(pl.len().alias("INSTANCES"))
6
7     df = df.with_columns(
8         pl.when(pl.col("GENE") == "vanR_gene_in_vanO_cluster"
9             ).then(pl.lit("vanRO"))
10            .otherwise(pl.col("GENE"))
11            ).alias("GENE")
12        )
13     df = df.with_columns(
14         pl.when(pl.col("GENE") == "vanS_gene_in_vanO_cluster"
15             ).then(pl.lit("vanSO"))
16            .otherwise(pl.col("GENE"))
17            ).alias("GENE")
18        )
19     df = df.with_columns(
20         pl.when(pl.col("GENE") == "Streptomyces_venezuelae Rox"
21             ).then(pl.lit("Rox-sv"))
22            .otherwise(pl.col("GENE"))
23            ).alias("GENE")
24        )
25     df = df.with_columns(
26         pl.when(pl.col("GENE") == "hsIB1/vipA/tssB"
27             ).then(pl.lit("hsIB1"))
28            .otherwise(pl.col("GENE"))
29            ).alias("GENE")
30        )
31     df = df.sort("INSTANCES", descending=True)
32     df = percenting_df(df)
33     return df

```

Fonte: Autoria própria.

Script 14 – Definição da função de criação do *dataframe* contendo a informação gênica de todos os arquivos importados.

```

1 def big_dataframe(df_list):
2     genes = pl.concat(df_list).unique()
3     genes = genes.drop("INSTANCES")
4     genes = genes.unique()
5     df_list = [genes.join(df, on="GENE", how="left")
6                 for df
7                 in df_list]
8     df_final = df_list[0]
9     for i in range(len(df_list) - 1):
10         df_final = df_final.join(
11             df_list[i + 1], how="full",
12             on="GENE", suffix=f"{i + 1}",
13             )
14     df = df_final.fill_null(0)
15     return df

```

Fonte: Autoria própria.

Script 15 – Função responsável por gerar as figuras ilustrando os ARGs e os VFs identificados.

```
1 def main_plot(df, path):
2     quantity = df.height
3     reports = df.width // 2
4
5     fig, ax = plt.subplots()
6     bottom = np.zeros(reports)
7
8     labels = df["GENE"].to_list()
9     samples = [f"SRR222782{i}" for i in range(25,33)]
10    colors = df["COLORS"].to_list()
11    for i in range(quantity):
12        y_values = df[i].select(pl.col(pl.Float64)).rows()[0]
13        ax.barh(samples, y_values, left=bottom,
14                color=colors[i],
15                )
16        bottom += y_values
17
18    ax.set_xlim(0, 100)
19    ax.set_xlabel("Abundancia (%)")
20    ax = creating_legend(ax, labels, colors)
21
22    fig = format_figure(fig)
23    fig.savefig(f"{path}")
24    return
```

Fonte: Autoria própria.

4 RESULTADOS E DISCUSSÕES

Nesta seção, os principais resultados obtidos durante o desenvolvimento deste trabalho são apresentados. Em particular, a Seção 4.1 aborda algumas características dos dados antes e após a etapa de filtragem. Já a Seção 4.2 apresenta os resultados provenientes da *pipeline* de caracterização taxonômica, evidenciando os principais gêneros identificados e discutindo seus papéis quando presentes no solo. Por sua vez, as Seções 4.3 e 4.4 ilustram alguns resultados intermediários sobre as estatísticas relacionadas as etapas de montagem e predição de genes. Por fim, as Seções 4.5 e 4.6 trazem resultados provenientes da execução da *pipeline* de caracterização do resistoma e do viruloma, juntamente com uma discussão sobre os principais aspectos observados a partir dos dados. Vale destacar que comparações com os resultados de Ordine et al. (2023) são apresentadas no Apêndice.

4.1 FILTRAGEM DOS DADOS

A Tabela 2 apresenta as métricas (extraídas do arquivo JSON), geradas pelo *software* **Fastp**, antes e depois da etapa de filtragem dos dados obtidos do sequenciamento. A análise da Tabela 2 revela que, inicialmente, os arquivos provenientes do sequenciamento apresentavam de 81 a 121 milhões de leituras *paired-end*, possuindo tamanho médio de 150 pares de base, com escores de qualidade Phred ≥ 30 (i.e., menos do que 1 erro por 1000 bases) (EWING et al., 1998), e conteúdo GC variando de 61.7% a 66.3%¹. Essas informações, quando somadas ao restante das métricas geradas pelo **Fastp**, atestam que os dados disponibilizados por Ordine et al. (2023) não apresentam características problemáticas relacionadas ao sequenciamento. Após a filtragem, embora a contagem total de leituras tenha diminuído ligeiramente, o conteúdo GC permanece similar aos valores pré-filtragem, indicando que as sequências removidas não concentravam um par de bases em específico. Além disso, o número de leituras filtradas é maior que o número total de sequências pós-filtragem, significando que a maior parte das leituras descartadas não reprovou pelos filtros de tamanho ou de qualidade mínima estabelecidos, mas sim por outros filtros como o de remoção de dados duplicados. Ainda, a Figura 3 ilustra o nível médio de qualidade *Phred* para cada amostra após a filtragem. Com base nessa figura, observa-se que esse índice se mantém acima de 35, o que implica uma confiabilidade de 99,97% em todas as posições das bases. Quando as leituras são comparadas, a *reverse read* apresenta um valor médio de qualidade ligeiramente menor que a *forward read*. Vale mencionar que esse comportamento é esperado, ocorrendo usualmente em dispositivos Illumina de sequenciamento *paired-end* (KWON et al., 2013).

¹ Conforme Lightfield, Fram e Ely (2011), o conteúdo GC no genoma de bactérias varia entre 16% e 75%; então, segundo Ismail (2023), valores extremos desse índice podem caracterizar um viés de sequenciamento.

Tabela 2 – Principais métricas das amostras antes e depois da etapa de controle de qualidade e limpeza.

		Amostras (códigos SRR)							
Parâmetros avaliados		22278225	22278226	22278227	22278228	22278229	22278230	22278231	22278232
Antes da filtragem	Total de leituras	83,1 M	81,0 M	114,0 M	104,6 M	116,6 M	93,1 M	121,0 M	101,7 M
	Bases Q30	11,8 Gb (94,5%)	11,4 Gb (94,1%)	16,0 Gb (93,7%)	14,8 Gb (94,1%)	16,5 Gb (94,1%)	13,2 Gb (94,2%)	17,1 Gb (94,1%)	14,3 Gb (93,8%)
	Conteúdo GC	63,3%	64,2%	62,8%	66,0%	65,3%	61,7%	66,3%	65,3%
	Tamanho médio	150, 150	150, 150	149, 149	149, 149	149, 149	149, 149	149, 149	149, 149
Depois da filtragem	Total de leituras	67,4 M	65,8 M	100,3 M	91,2 M	102,0 M	82,0 M	105,9 M	89,6 M
	Bases Q30	9,6 Gb (95,0%)	9,3 Gb (94,6%)	14,0 Gb (93,4%)	12,8 Gb (93,8%)	14,4 Gb (93,9%)	11,6 Gb (94,0%)	14,9 Gb (93,8%)	12,6 Gb (93,6%)
	Conteúdo GC	63,3%	64,2%	62,7%	66,0%	65,2%	61,5%	66,2%	65,2%
	Tamanho médio	149, 149	149, 149	149, 149	149, 149	149, 149	149, 149	149, 149	149, 149
Leituras filtradas		81,1 M (97,642%)	79,0 M (97,496%)	114,0 M (99,990%)	104,6 M (99,989%)	116,6 M (99,992%)	93,1 M (99,990%)	121,0 M (99,991%)	101,7 M (99,991%)

Fonte: Autoria própria.

4.2 ANÁLISE TAXONÔMICA

A análise de abundância de gêneros (Figura 4) revelou a presença predominante de *Streptomyces*, *Bradyrhizobium*, *Nocardioides*, *Mycolicibacterium*, *Micromonospora*, *Pseudomonas*, *Sphingomonas* e *Mycobacterium*. Esses gêneros desempenham papéis cruciais nos ecossistemas do solo, contribuindo para a saúde e fertilidade do ambiente agrícola. *Streptomyces* e *Micromonospora* são, comumente, reconhecidos por sua capacidade de produzir compostos bioativos, como antibióticos (KONWAR et al., 2024; SOKOŁOWSKI et al., 2024), e por sua atuação na decomposição de matéria orgânica e degradação de polímeros (SENKO et al., 2024). *Bradyrhizobium* é um dos gêneros bacterianos mais conhecidos na agricultura, sendo responsável pela fixação de nitrogênio atmosférico, formando simbiose com leguminosas e promovendo a disponibilidade desse nutriente essencial às plantas (SZPUNAR-KROK et al., 2023; SARAO et al., 2024). *Nocardioides* e *Sphingomonas* são associados à degradação de compostos orgânicos complexos, importantes para o ciclo de carbono (ZHANG et al., 2024). O estudo conduzido por Zhang et al. (2024) identificou *Nocardioides* como um dos principais gêneros bacterianos responsáveis pela modificação de genes de decomposição de carbono no solo. *Pseudomonas* e *Mycolicibacterium* são amplamente estudados por suas funções na promoção do crescimento vegetal (YANG et al., 2021; RAI et al., 2024), detoxificação de metais pesados (MADHOGARIA et al., 2024) e na proteção contra fitopatógenos (KHATRI et al., 2024). Ainda, *Mycobacterium* inclui

espécies que participam de processos de degradação de matéria orgânica no solo (MON et al., 2024). Todos os principais gêneros encontrados aqui estão amplamente distribuídos em solos e desempenham importantes funções na ciclagem de nutrientes e na manutenção da saúde do ambiente.

4.3 ESTATÍSTICAS RELATIVAS À MONTAGEM DE CONTIGS

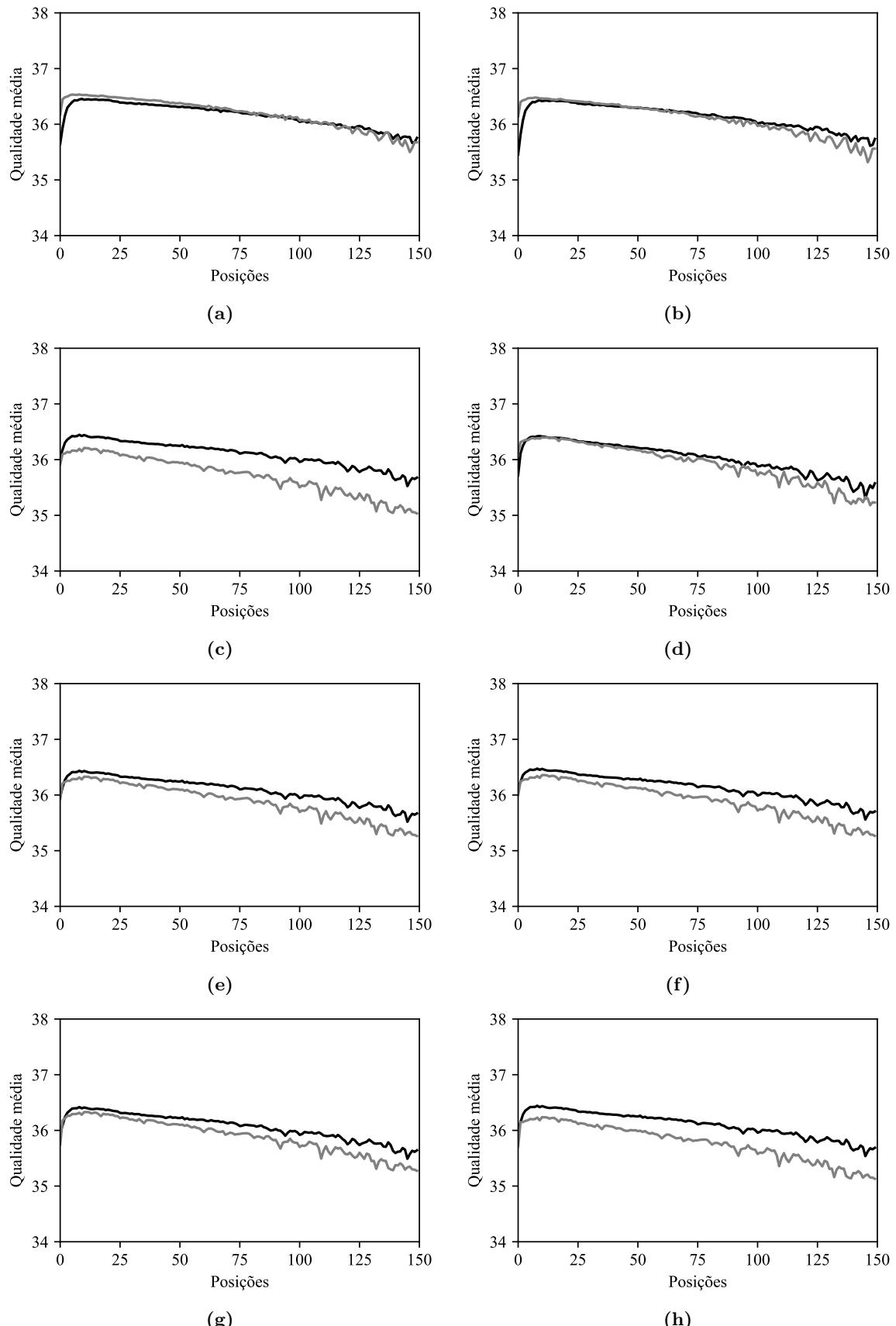
A Tabela 3 traz algumas das características da montagem realizada pelo MEGAHIT. Das 8 amostras analisadas, o número total de *contigs* varia entre 107,0 mil e 190,3 mil, tendo os maiores *contigs* tamanhos entre 42451 e 454752 pares de bases. Acerca de outras métricas apresentadas, o N50 é definido como o maior comprimento que contém 50% dos nucleotídeos presentes nos *contigs* (JGI, 2001), enquanto o L50 é definido como o número mínimo de *contigs*, ordenados do maior para o menor, cuja soma dos comprimentos representa pelo menos 50% do total de bases da montagem (JAYAKUMAR; SAKAKIBARA, 2019). A contiguidade da montagem dos dados está relacionada com essas duas estatísticas; em resumo, quanto maior o valor de N50 e menor o valor de L50, tem-se que uma quantidade reduzida de *contigs* abrange 50% do total de nucleotídeos, sugerindo assim uma montagem mais completa dos dados (JAYAKUMAR; SAKAKIBARA, 2019). Outras métricas, como o número de *gaps* ou bases desconhecidas (N), não foram apresentadas, uma vez que seus valores foram zero em todas as amostras.

Tabela 3 – Características dos *contigs* montados pelo software MEGAHIT.

Parâmetros avaliados	Amostras (códigos SRR)							
	22278225	22278226	22278227	22278228	22278229	22278230	22278231	22278232
Maior <i>contig</i>	233035	84108	454752	42451	42495	203756	105590	81105
Média	581,93	556,80	554,85	600,76	538,81	652,89	616,28	548,73
N50	586	554	551	624	546	679	633	545
L50	394470 (29,9%)	330300 (30,9%)	433100 (30,1%)	504130 (27,7%)	529550 (32,6%)	485330 (25,5%)	474630 (27,2%)	415410 (30,9%)
Total de <i>contigs</i>	132,9 K	107,0 K	144,1 K	181,9 K	162,5 K	190,3 K	174,3 K	134,4 K
Total de bases	76,7 M	59,6 M	79,9 M	109,2 M	87,5 M	124,2 M	107,4 M	73,8 M

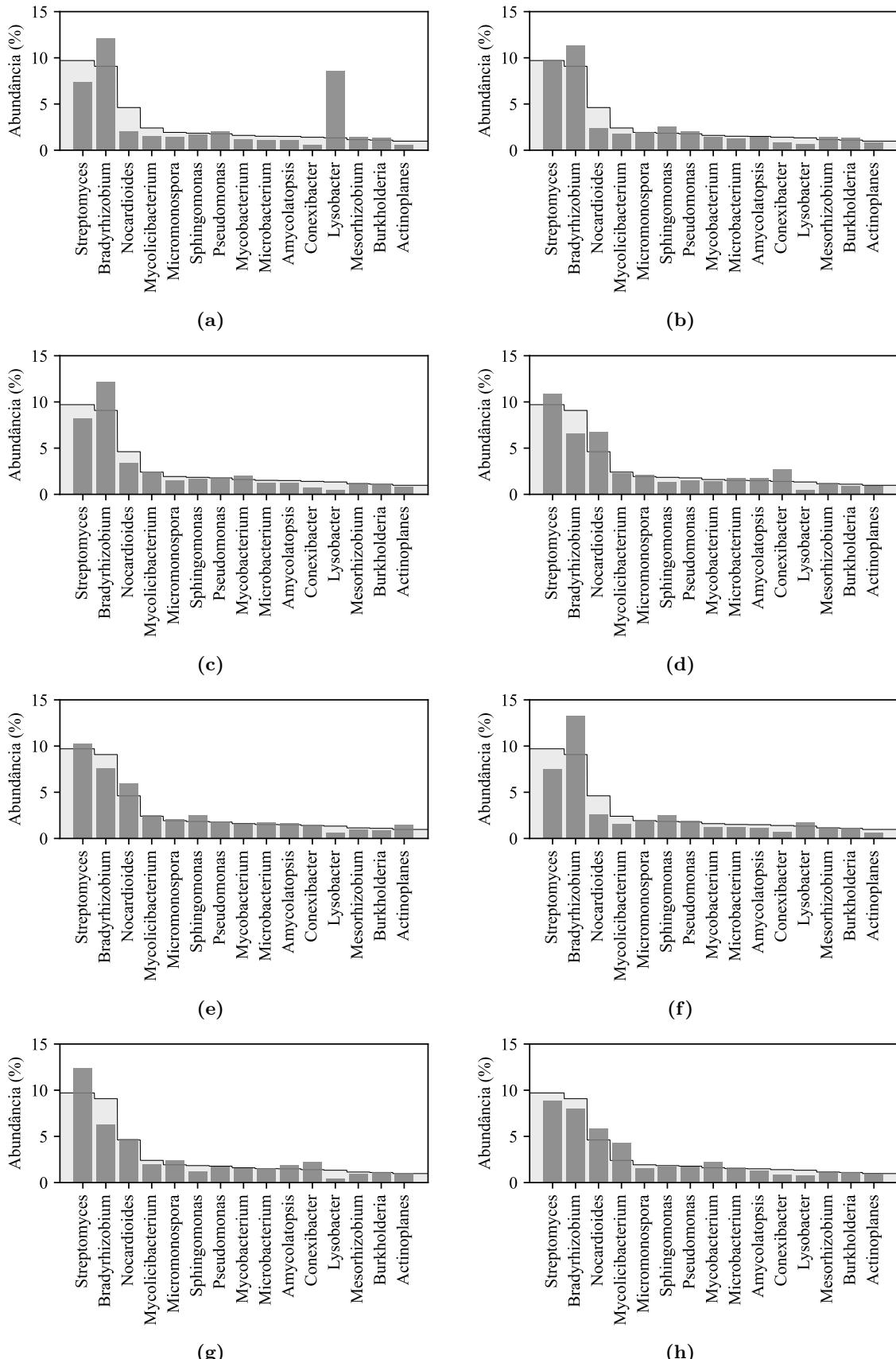
Fonte: Autoria própria.

Figura 3 – Qualidade média das bases após a etapa de controle de qualidade e limpeza para a *forward read* (linha escura) e a *reverse read* (linha clara). (a) SRR22278225. (b) SRR22278226. (c) SRR22278227. (d) SRR22278228. (e) SRR22278229. (f) SRR22278230. (g) SRR22278231. (h) SRR22278232.



Fonte: Autoria própria.

Figura 4 – Análise taxonômica considerando a abundância relativa dos gêneros presentes na amostra (barras cinzas) e a média do conjunto (linha escura sólida). (a) SRR22278225. (b) SRR22278226. (c) SRR22278227. (d) SRR22278228. (e) SRR22278229. (f) SRR22278230. (g) SRR22278231. (h) SRR22278232.



Fonte: Autoria própria.

4.4 RESULTADOS DECORRENTES DA PREDIÇÃO DE GENES

A Tabela 4 apresenta os resultados obtidos a partir do arquivo .gff gerado na etapa de predição de genes. O número total de regiões de codificação (CDS)² preditas varia entre 521,3 mil e 1,1 milhão dentre as amostras. Dentre esses CDS, aproximadamente de 75% a 79% correspondem a proteínas hipotéticas, i.e., proteínas que provavelmente são traduzidas, mas que não possuem evidência experimental de tradução (DESLER; DURHUUS; RASMUSSEN, 2012). Também, foram preditos entre 490 e 1220 genes codificantes de proteínas de resistência a múltiplas drogas (MDRs), associadas à presença de ARGs nas comunidades microbianas (NIKAIDO, 2009), além de alguns ARGs específicos (e.g., *sbmA* e *arpA*). Ainda, identificou-se uma categoria de genes responsável pela codificação de proteínas de virulência, um termo que define um conjunto de proteínas que fazem parte dos mecanismos de VFs de alguma bactéria.

Tabela 4 – Resultados decorrentes da etapa de predição de genes.

Parâmetros avaliados	Amostras (códigos SRR)							
	22278225	22278226	22278227	22278228	22278229	22278230	22278231	22278232
CDS preditos	684,4 K	521,3 K	695,1 K	982,4 K	761,0 K	1,1 M	960,5 K	634,3 K
Proteínas Hipotéticas	539,6 K (78,84%)	409,2 K (78,51%)	545,6 K (78,49%)	739,4 K (75,26%)	585,6 K (76,95%)	882,3 K (78,02%)	718,6 K (74,82%)	482,4 K (76,05%)
MDRs e ARGs	790	543	647	638	475	1220	794	495
Proteínas de virulência	34	20	23	27	38	33	38	20

Fonte: Autoria própria.

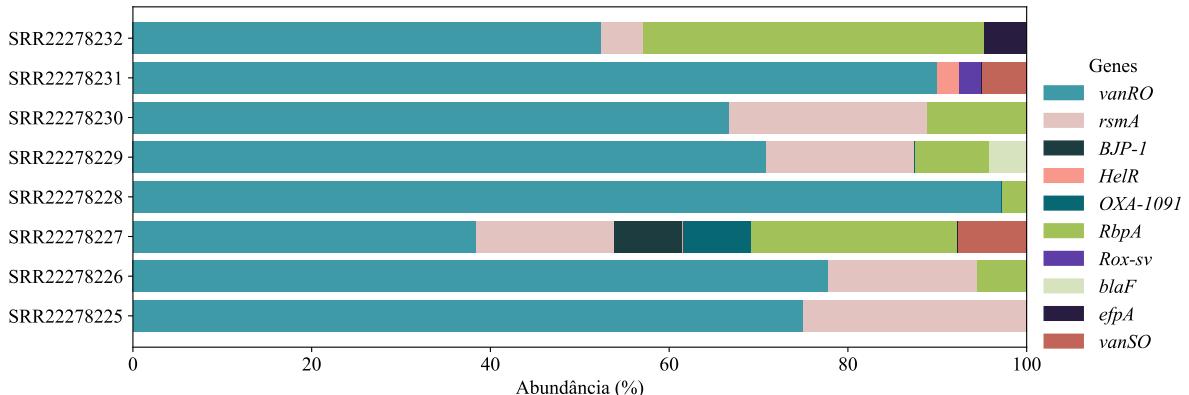
4.5 CARACTERIZAÇÃO DO RESISTOMA

A Figura 5 demonstra os genes que compõem o resistoma das amostras consideradas. Desses genes³, alguns conferem resistências variadas a antibióticos, como o *vanRO*. Os genes *vanR* e *vanO* expressam proteínas que promovem a alteração de alvos de glicopeptídeos presentes na parede celular bacteriana (STOGIOS; SAVCHENKO, 2020). Esses genes estão presentes em bactérias Gram positivas, como *Staphylococcus* e *Enterococcus*. Todavia, essas bactérias não foram detectadas nos 8 solos analisados aqui (veja a Figura 4), indicando que tais bactérias estão presentes em pequenas quantidades no solo. O gene *rsmA*, cujo produto regula um sistema de bombas de efluxo, promove a expulsão de antibióticos como o fenicol, a diaminopirimidina e a fluoroquinolona para fora da

² Conforme UniProtKB (2024), uma região de codificação (CDS) determina um segmento de DNA ou RNA responsável por conter uma sequência de aminoácidos de uma proteína.

³ Embora algumas funções gênicas sejam discutidas, destaca-se que nenhuma pipeline de identificação funcional foi implementada. As informações apresentadas foram coletadas diretamente dos bancos de dados CARD e VFDB.

Figura 5 – Genes de resistência a antibióticos identificados nas diferentes amostras analisadas.



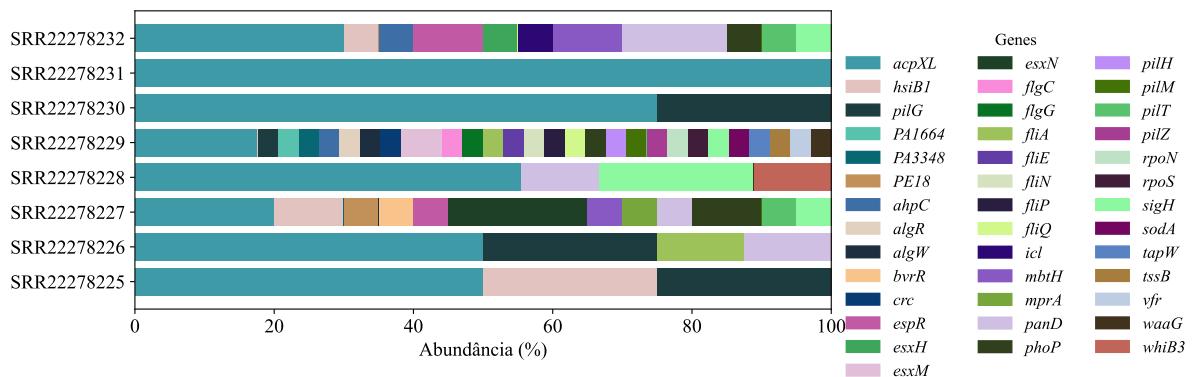
Fonte: Autoria própria.

célula bacteriana (BRENCIC; LORY, 2009). Esse gene pode ser encontrado em *Pseudomonas*, bactéria que foi detectada em todos os solos analisados (Figura 4). Dentre as espécies de *Pseudomonas*, a espécie *P. aeruginosa* é a que está associada a doenças em humanos. Tal microrganismo apresenta ainda resistência intrínseca e adquirida a vários antibióticos devido a alterações na permeabilidade da membrana externa e expressão de bombas de efluxo. O *RbpA* é um regulador global de RNA polimerase, que indiretamente, promove a expressão de genes que facilitam a resistência à rifampicina (HU et al., 2012). Rifampicina é um antibiótico muito utilizado no tratamento de infecções causadas por *Mycobacterium tuberculosis*, como a tuberculose. *Mycobacterium* foi detectada em todos os solos analisados aqui (Figura 4).

4.6 CARACTERIZAÇÃO DO VIRULOMA

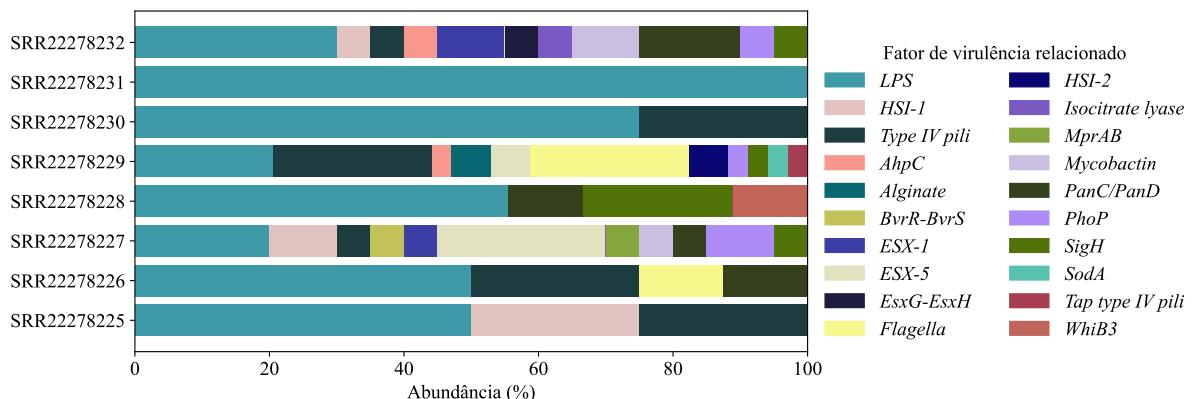
O VFDB (CHEN et al., 2005) apresenta, conjuntamente à nomenclatura dos genes, informações relacionadas com as funções gênicas, com a cepa da bactéria cujo material genético foi alinhado no momento da classificação e com os fatores de virulência relacionados. Dessa maneira, através da Figura 6, percebe-se que os resultados apresentaram maior variabilidade do que foi encontrado no caso dos ARGs. Ao analisar os resultados obtidos, pode-se inferir que o gene *acpXL* foi o único identificado em todas as amostras analisadas, sendo especialmente presente nos solos SRR22278230 e SRR22278231. Esse gene está relacionado à biosíntese de Lipopolissacárido (LPS), sendo uma molécula exclusiva da parede de bactérias Gram negativas. A presença de LPS em todos os solos, com 100% de presença no solo SRR22278231, conforme se observa na Figura 7, é um fato interessante. Por ser um solo de criação de gado, esse LPS deve indicar contaminação do solo com bactérias intestinais oriundas de fezes do próprio gado bovino. Outras amostras de solo demonstraram maior variabilidade gênica, sem apresentar destaque para um gene em específico, conforme ocorreu em SRR22278227 e SRR22278229. Das funções apresen-

Figura 6 – Genes de fatores de virulência identificados nas diferentes amostras analisadas.



Fonte: Autoria própria.

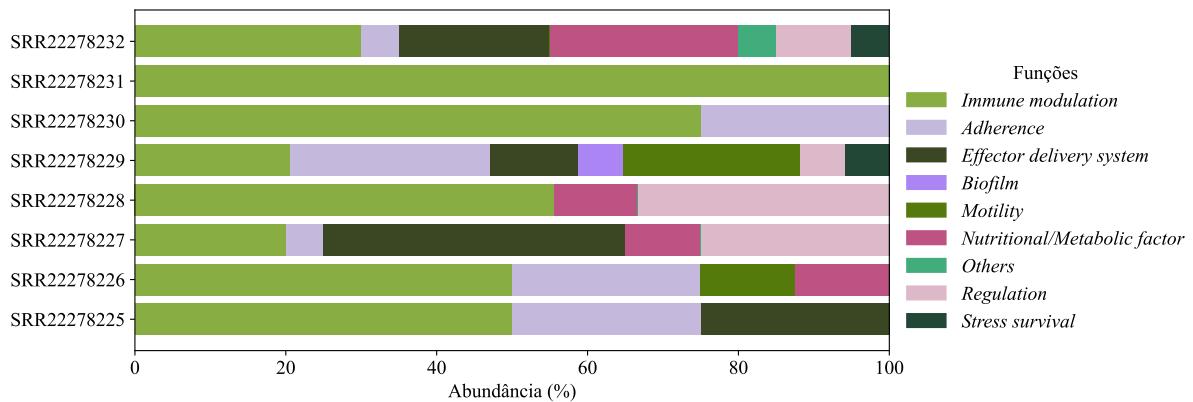
Figura 7 – Fatores de virulência relacionados aos genes identificados nas diferentes amostras.



Fonte: Autoria própria.

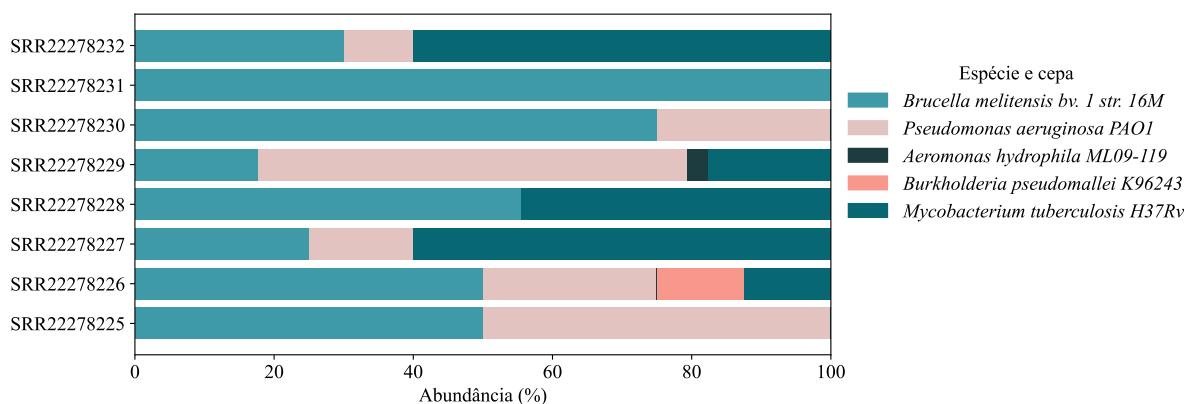
tadas pelo banco de dados, conforme a Figura 8, modulação imune e aderência compõem uma parcela significativa do que foi observado, com outros mecanismos relacionados com fatores nutricionais/metabólicos ou de motilidade se mostrando presentes. Sobre as cepas das quais os genes foram identificados, apresentadas na Figura 9, a *Brucella melitensis* bv. 1 str. 16M se apresentou em todas as amostras, enquanto as cepas *Burkholderia pseudomallei* K96243 e *Aeromonas hydrophila* ML09-119 foram observadas apenas em uma amostra cada.

Figura 8 – Funções associadas aos genes de fatores de virulência identificados nas diferentes amostras analisadas.



Fonte: Autoria própria.

Figura 9 – Cepas bacterianas cujo material genético foi alinhado aos genes de fatores de virulência identificados nas diferentes amostras analisadas.



Fonte: Autoria própria.

5 CONSIDERAÇÕES FINAIS

Nesta seção, as conclusões deste trabalho são discutidas, abordando brevemente o que foi realizado e os principais resultados obtidos. Em seguida, algumas sugestões para trabalhos futuros são apresentadas, visando dar continuidade ao presente trabalho. Por fim, uma lista com os trabalhos publicados e/ou em vias de publicação é fornecida.

5.1 CONCLUSÕES

Neste trabalho, duas *pipelines* foram implementadas visando a análise de dados de metagenômica provenientes do sequenciamento do tipo *shotgun* de amostras de solo, disponibilizados publicamente por Ordine et al. (2023) através do *sequence read archive* (SRA) sob o *BioProject PRJNA900430* (NCBI, 2024a). Em particular, a primeira *pipeline* implementada, desenvolvida com base no protocolo descrito por (LU et al., 2022), visa realizar a classificação taxonômica das comunidades microbianas presentes nas amostras, utilizando para isso os *softwares* *Fastp* (CHEN, 2023), *Kraken2* (WOOD; LU; LANGMEAD, 2019) e *Bracken* (LU et al., 2017). Com a execução dessa *pipeline*, foram caracterizados gêneros bacterianos importantes (e.g., *Bradyrhizobium*, *Nocardioides* e *Pseudomonas*) para a saúde dos organismos que compõem os ecossistemas presentes no solo. A segunda *pipeline*, implementada com base em (ORDINE et al., 2023), visa caracterizar o resistoma e o viruloma através da identificação de genes de resistência a antibióticos (ARGs) e de fatores de virulência (VFs), respectivamente, empregando para isso os *softwares* *Fastp*, *MEGAHIT* (LI et al., 2015), *Prokka* (SEEMANN, 2014) e *ABRicate* (SEEMANN, 2020) associado aos bancos de dados *CARD* (MCARTHUR et al., 2013) e *VFDB* (CHEN et al., 2005). A execução dessa *pipeline* permitiu identificar importantes genes de resistência a antibióticos (e.g., *vanRO*, *rsmA* e *RbpA*) e de fatores de virulência (e.g., *acpXL*, *hsbB1* e *pilG*), destacando assim a necessidade de monitoramento continuado para fins epidemiológicos. Ainda, discussões sobre alguns outros resultados intermediários, tais como as estatísticas de controle de qualidade, de montagem e de predição de gene, foram conduzidas para atestar a acurácia das análises. Vale destacar que, visando possibilitar a reprodutibilidade dos resultados apresentados aqui, os *scripts* (implementados em linguagem **GNU Bash**) para o processamento das amostras (dados brutos), assim como os códigos (desenvolvidos em **Python**) para realizar a análise dos resultados foram disponibilizados publicamente (devidamente documentados).

5.2 SUGESTÕES DE TRABALHOS FUTUROS

Como ideias para o desenvolvimento de trabalhos futuros na área, sugere-se:

- Realizar uma análise de diversidade α e β sobre os dados taxonômicos;
- Utilizar outros bancos de dados para a caracterização do resistoma, introduzindo um catálogo gênico mais completo para a identificação de genes;
- Implementar uma *pipeline* de análise funcional, visando compreender de forma mais completa o papel de alguns genes e suas participações em vias metabólicas; e
- Realizar análises estatísticas pertinentes sobre os resultados obtidos, avaliando em como os resultados se relacionam com o perfil do solo apresentado.

5.3 TRABALHOS PUBLICADOS

Durante o desenvolvimento deste trabalho de pesquisa, os seguintes materiais foram publicados e/ou estão em vias de publicação:

- LEANDRO, Roberto Marafon *et al.* CONSIDERAÇÕES SOBRE A ANÁLISE DE DADOS DE METAGENÔMICA OBTIDOS DE AMOSTRAS DE SOLO. In: Anais do XIV Seminário de Extensão e Inovação & XXIX Seminário de Iniciação Científica e Tecnológica da UTFPR (SEI/SICITE), Francisco Beltrão, PR, 2024. Disponível em: <<https://doi.org/10.29327/seisicite2024.967380>>.
- LEANDRO, Roberto Marafon *et al.* ANALYSIS OF METAGENOMIC DATA OBTAINED FROM SOIL SAMPLES: TAXONOMIC, FUNCTIONAL AND GENIC CHARACTERIZATION. In: Journal of Microbiological Methods, 2025 (*em redação*).

REFERÊNCIAS

- ABRAM, F. Systems-based approaches to unravel multispecies microbial community functioning. **Computational and Structural Biotechnology Journal**, v. 13, p. 24–32, 2015.
- AGROADVANCE. **6 maiores produtores de soja do mundo: quando e quanto produzem?** 2024. Disponível em: <<https://agroadvance.com.br/blog-6-maiores-produtores-de-soja-do-mundo/>>. Acesso em: 04 mar. 2025.
- ALCOCK, B. P. et al. Card 2023: expanded curation, support for machine learning, and resistome prediction at the comprehensive antibiotic resistance database. **Nucleic acids research**, Oxford University Press, v. 51, n. D1, p. D690–D699, 2023.
- ALTSCHUL, S. F. et al. Basic local alignment search tool. **Journal of molecular biology**, Elsevier, v. 215, n. 3, p. 403–410, 1990.
- ANDREWS, S. **FastQC: A Quality Control Tool for High Throughput Sequence Data.** 2010. Disponível em: <<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>>. Acesso em: 04 mar. 2025.
- ARAGÃO, A.; CONTINI, E. **O AGRO NO BRASIL E NO MUNDO: UMA SÍNTSE DO PERÍODO DE 2000 A 2020.** 2023. Disponível em: <<https://www.embrapa.br/documents/10180/62618376/O+AGRO+NO+BRASIL+E+NO+MUNDO.pdf>>. Acesso em: 04 mar. 2025.
- AYLING, M.; CLARK, M. D.; LEGGETT, R. M. New approaches for metagenome assembly with short reads. **Briefings in Bioinformatics**, p. 1–13, 2019.
- BANERJEE, S.; HEIJDEN, M. G. V. D. Soil microbiomes and one health. **Nature Reviews Microbiology**, Nature Publishing Group UK London, v. 21, n. 1, p. 6–20, 2023.
- BESEMER, J.; LOMSADZE, A.; BORODOVSKY, M. Genemarks: a self-training method for prediction of gene starts in microbial genomes. implications for finding sequence motifs in regulatory regions. **Nucleic acids research**, Oxford University Press, v. 29, n. 12, p. 2607–2618, 2001.
- BHARTI, R.; GRIMM, D. G. Current challenges and best-practice protocols for microbiome analysis. **Briefings in Bioinformatics**, v. 22, n. 1, p. 178–193, Jan. 2021.
- BOISVERT, S. et al. Ray meta: scalable de novo metagenome assembly and profiling. **Genome Biology**, v. 13, n. 12, p. R122, 2012.
- BOLGER, A. M.; LOHSE, M.; USADEL, B. Trimmomatic: a flexible trimmer for illumina sequence data. **Bioinformatics**, Oxford University Press, v. 30, n. 15, p. 2114–2120, 2014.
- BORODOVSKY, M. et al. Detection of new genes in a bacterial genome using markov models for three gene classes. **Nucleic acids research**, Oxford University Press, v. 23, n. 17, p. 3554–3562, 1995.

- BORTOLAIA, V. et al. Resfinder 4.0 for predictions of phenotypes from genotypes. **Journal of Antimicrobial Chemotherapy**, Oxford University Press, v. 75, n. 12, p. 3491–3500, 2020.
- BREITWIESER, F. P.; SALZBERG, S. L. Pavian: interactive analysis of metagenomics data for microbiome studies and pathogen identification. **Bioinformatics**, Oxford University Press, v. 36, n. 4, p. 1303–1304, 2020.
- BRENCIC, A.; LORY, S. Determination of the regulon and identification of novel mrna targets of pseudomonas aeruginosa rsma. **Molecular microbiology**, Wiley Online Library, v. 72, n. 3, p. 612–632, 2009.
- BROWN, J.; PIRRUNG, M.; MCCUE, L. A. FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. **Bioinformatics**, Oxford University Press, v. 33, n. 19, p. 3137–3139, 2017.
- BUENO, A. D. F. et al. Challenges for adoption of integrated pest management (ipm): the soybean example. **Neotropical Entomology**, v. 50, p. 5–20, 2021.
- CHEN, L. et al. VFDB: a reference database for bacterial virulence factors. **Nucleic acids research**, Oxford University Press, v. 33, n. suppl_1, p. D325–D328, 2005.
- CHEN, S. Ultrafast one-pass fastq data preprocessing, quality control, and deduplication using fastp. **Imeta**, Wiley Online Library, v. 2, n. 2, p. e107, 2023.
- CLAUSEN, P. T.; AARESTRUP, F. M.; LUND, O. Rapid and precise alignment of raw reads against redundant databases with kma. **BMC bioinformatics**, Springer, v. 19, p. 1–8, 2018.
- COMPANT, S. et al. Harnessing the plant microbiome for sustainable crop production. **Nature Reviews Microbiology**, v. 22, n. 8, p. 567–584, August 2024.
- COMPEAU, P. E.; PEVZNER, P. A.; TESLER, G. How to apply de bruijn graphs to genome assembly. **Nature Biotechnology**, v. 29, n. 11, p. 987–991, 2011.
- CONFEDERAÇÃO DA AGRICULTURA E PECUÁRIA DO BRASIL. **Panorama do Agro**. 2024. Disponível em: <<https://www.cnabrasil.org.br/cna/panorama-do-agro>>. Acesso em: 04 mar. 2025.
- DELCHER, A. L. et al. Identifying bacterial genes and endosymbiont dna with glimmer. **Bioinformatics**, Oxford University Press, v. 23, n. 6, p. 673–679, 2007.
- DESLER, C.; DURHUUS, J. A.; RASMUSSEN, L. J. Genome-wide screens for expressed hypothetical proteins. **Functional Genomics: Methods and Protocols**, Springer, p. 25–38, 2012.
- DRÖGE, J.; GREGOR, I.; MCHARDY, A. C. Taxator-tk: precise taxonomic assignment of metagenomes by fast approximation of evolutionary neighborhoods. **Bioinformatics**, Oxford University Press, v. 31, n. 6, p. 817–824, 2015.
- EDWIN, N. R. et al. An in-depth evaluation of metagenomic classifiers for soil microbiomes. **Environmental Microbiome**, Springer, v. 19, n. 1, p. 19, 2024.

- EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA. **TRAJETÓRIA DA AGRICULTURA BRASILEIRA**. 2022. Disponível em: <<https://www.embrapa.br/vsiao/trajetoria-da-agricultura-brasileira>>. Acesso em: 04 mar. 2025.
- EWELS, P. et al. MultiQC: summarize analysis results for multiple tools and samples in a single report. **Bioinformatics**, Oxford University Press, v. 32, n. 19, p. 3047–3048, 2016.
- EWING, B. et al. Base-calling of automated sequencer traces using phred. ii. error probabilities. **Genome Research**, v. 8, n. 3, p. 186–194, 1998. Disponível em: <<https://genome.cshlp.org/content/8/3/186>>.
- FAO. **Soil Management and Conservation for Small Farms: Strategies and Methods of Introduction, Technologies and Equipment**. Rome: Food and Agriculture Organization of the United Nations, 2017.
- FAO, IFAD, UNICEF, WFP e WHO. **The State of Food Security and Nutrition in the World 2024 – Financing to end hunger, food insecurity and malnutrition in all its forms**. Rome: FAO, 2024.
- FELDGARDEN, M. et al. Amrfinderplus and the reference gene catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. **Scientific reports**, Nature Publishing Group UK London, v. 11, n. 1, p. 12728, 2021.
- FLORENSA, A. F. et al. Resfinder—an open online resource for identification of antimicrobial resistance genes in next-generation sequencing data and prediction of phenotypes from genotypes. **Microbial Genomics**, Microbiology Society, v. 8, n. 1, p. 000748, 2022.
- FLYGARE, S. et al. Taxonomer: an interactive metagenomics analysis portal for universal pathogen detection and host mrna expression profiling. **Genome Biology**, BioMed Central, v. 17, n. 1, p. 111, 2016.
- FUENTES-LLANILLO, R. et al. Expansion of no-tillage practice in conservation agriculture in brazil. **Soil and Tillage Research**, v. 208, p. 104877, 2021.
- GEMAYEL, K.; LOMSADZE, A.; BORODOVSKY, M. Metagenemark-2: improved gene prediction in metagenomes. **BioRxiv**, Cold Spring Harbor Laboratory, 2022.
- GNU, P. **Free Software Foundation. Bash (3.2. 48)[Unix shell program]**. 2007.
- GREEN, M. R.; SAMBROOK, J. **Molecular Cloning: A Laboratory Manual**. 4th edition. ed. [S.l.]: Cold Spring Harbor Laboratory Press, 2012.
- HAIDER, B. et al. Omega: an overlap-graph de novo assembler for metagenomics. **Bioinformatics**, v. 30, n. 19, p. 2717–2722, October 2014.
- HOWE, A. C. et al. Tackling soil diversity with the assembly of large, complex metagenomes. **Proceedings of the National Academy of Sciences of the United States of America**, v. 111, n. 13, p. 4904–4909, 2014.
- HU, T. et al. Next-generation sequencing technologies: An overview. **Human Immunology**, v. 82, n. 11, p. 801–811, 2021. ISSN 0198-8859. Next Generation Sequencing and its Application to Medical Laboratory Immunology. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0198885921000628>>.

- HU, Y. et al. Mycobacterium tuberculosis rbpa protein is a new type of transcriptional activator that stabilizes the σ a-containing rna polymerase holoenzyme. **Nucleic acids research**, Oxford University Press, v. 40, n. 14, p. 6547–6557, 2012.
- HUSON, D. H. et al. Megan analysis of metagenomic data. **Genome research**, Cold Spring Harbor Lab, v. 17, n. 3, p. 377–386, 2007.
- HUSON, D. H. et al. Megan community edition - interactive exploration and analysis of large-scale microbiome sequencing data. **PLoS Computational Biology**, Public Library of Science, v. 12, n. 6, p. e1004957, 2016.
- HYATT, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. **BMC bioinformatics**, Springer, v. 11, p. 1–11, 2010.
- ILLUMINA. **Experience Faster Library Prep**. 2024. Disponível em: <<https://www.illumina.com/techniques/sequencing/ngs-library-prep.html>>. Acesso em: 04 mar. 2025.
- ILLUMINA. **FastQ Files**. 2024. Disponível em: <<https://help.basespace.illumina.com/files-used-by-basespace/fastq-files>>. Acesso em: 04 mar. 2025.
- ILLUMINA. **Sequence file formats for a variety of data analysis options**. 2024. Disponível em: <<https://www.illumina.com/informatics/sequencing-data-analysis/sequence-file-formats.html>>. Acesso em: 04 mar. 2025.
- ILLUMINA. **Sequencing platforms**. 2024. Disponível em: <<https://www.illumina.com/systems/sequencing-platforms.html>>. Acesso em: 04 mar. 2025.
- ILLUMINA. **What is NGS?** 2024. Disponível em: <<https://www.illumina.com/science/technology/next-generation-sequencing.html>>. Acesso em: 04 mar. 2025.
- ISMAIL, H. D. **Bioinformatics: A Practical Guide to Next Generation Sequencing Data Analysis**. [S.l.]: Chapman and Hall/CRC, 2023.
- JAYAKUMAR, V.; SAKAKIBARA, Y. Comprehensive evaluation of non-hybrid genome assembly tools for third-generation pacbio long-read sequence data. **Briefings in bioinformatics**, Oxford University Press, v. 20, n. 3, p. 866–876, 2019.
- JOSHI, N.; FASS, J. **Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files**. 2011. Disponível em: <<https://github.com/najoshi/sickle>>. Acesso em: 04 mar. 2025.
- KHATRI, S. et al. Pseudomonas is a key player in conferring disease suppressiveness in organic farming. **Plant and Soil**, Springer, v. 503, n. 1, p. 85–104, 2024.
- KIM, D. et al. Centrifuge: rapid and sensitive classification of metagenomic sequences. **Genome Research**, Cold Spring Harbor Laboratory Press, v. 26, n. 12, p. 1721–1729, 2016.
- KONWAR, A. N. et al. Antimicrobial potential of Streptomyces sp. NP73 isolated from the forest soil of northeast india against multi-drug resistant escherichia coli. **Letters in Applied Microbiology**, Oxford University Press, v. 77, n. 9, p. ova086, 2024.
- KRAKENTOOLS DEVELOPERS. **KrakenTools: Utilities for working with Kraken2 output files**. 2021. Disponível em: <<https://github.com/jenniferlu717/KrakenTools>>. Acesso em: 04 mar. 2025.

- KWON, S. et al. In-depth analysis of interrelation between quality scores and real errors in illumina reads. In: IEEE. **2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC)**. [S.l.], 2013. p. 635–638.
- LANGMEAD, B.; SALZBERG, S. L. Fast gapped-read alignment with bowtie 2. **Nature Methods**, Nature Publishing Group, v. 9, n. 4, p. 357–359, 2012.
- LI, D. et al. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. **Bioinformatics**, v. 31, n. 10, p. 1674–1676, May 2015. Disponível em: <<https://doi.org/10.1093/bioinformatics/btv033>>.
- LI, H. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. **arXiv preprint arXiv:1303.3997**, 2013.
- LIGHTFIELD, J.; FRAM, N. R.; ELY, B. Across bacterial phyla, distantly-related genomes with similar genomic gc content have similar patterns of amino acid usage. **PloS one**, Public Library of Science San Francisco, USA, v. 6, n. 3, p. e17677, 2011.
- LIU, B.; POP, M. Ardb—antibiotic resistance genes database. **Nucleic acids research**, Oxford University Press, v. 37, n. suppl_1, p. D443–D447, 2009.
- LIU, Y.-X. et al. A practical guide to amplicon and metagenomic analysis of microbiome data. **Protein & cell**, Oxford University Press, v. 12, n. 5, p. 315–330, 2021.
- LOBANOV, V.; GOBET, A.; JOYCE, A. Ecosystem-specific microbiota and microbiome databases in the era of big data. **Environmental Microbiome**, Springer, v. 17, n. 1, p. 37, 2022.
- LU, J. et al. Bracken: estimating species abundance in metagenomics data. **PeerJ Computer Science**, PeerJ Inc., v. 3, p. e104, 2017.
- LU, J. et al. Metagenome analysis using the kraken software suite. **Nature Protocols**, Nature Publishing Group, v. 17, n. 12, p. 2815–2839, 2022.
- LUO, R. et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. **GigaScience**, v. 1, p. 18, 2012.
- MADHOGARIA, B. et al. Alleviation of heavy metals chromium, cadmium and lead and plant growth promotion in vigna radiata l. plant using isolated pseudomonas geniculata. **International Microbiology**, Springer, p. 1–17, 2024.
- MARTIN, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. **EMBnet.journal**, Cold Spring Harbor Laboratory Press, v. 17, n. 1, p. 10–12, 2011.
- MCARTHUR, A. G. et al. The comprehensive antibiotic resistance database. **Antimicrobial Agents and Chemotherapy**, v. 57, n. 7, p. 3348–3357, 2013.
- MENZEL, P.; NG, K. L.; KROGH, A. Kaiju: fast and sensitive taxonomic classification for metagenomics. **Nature Communications**, Nature Publishing Group, v. 7, p. 11257, 2016.
- MINISTÉRIO DA AGRICULTURA E DA PECUÁRIA. Nações Unidas adotam Declaração Política sobre Resistência Antimicrobiana durante a AGNU. [S.l.], 2024.

- MON, M. L. et al. Exploring the cellulolytic activity of environmental mycobacteria. **Tuberculosis**, Elsevier, v. 147, p. 102516, 2024.
- NAMIKI, T. et al. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. **Nucleic Acids Research**, v. 40, n. 20, p. e155, 2012.
- NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION. **National Library of Medicine: National Center for Biotechnology Information**. 2024. Disponível em: <<https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA900430>>. Acesso em: 04 mar. 2025.
- NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION. **SRA Toolkit**. 2024. Disponível em: <<https://github.com/ncbi/sra-tools>>. Acesso em: 04 mar. 2025.
- NIKAIKO, H. Multidrug resistance in bacteria. **Annual review of biochemistry**, Annual Reviews, v. 78, n. 1, p. 119–146, 2009.
- NURK, S. et al. metaSPAdes: a new versatile metagenomic assembler. **Genome Research**, v. 27, n. 5, p. 824–834, 2017.
- OND OV, B. D.; BERGMAN, N. H.; PHILLIPPY, A. M. An interactive metagenomic visualization tool. **BMC Bioinformatics**, BioMed Central, v. 12, n. 1, p. 385, 2011.
- ORDINE, J. V. W. et al. Metagenomic insights for antimicrobial resistance surveillance in soils with different land uses in brazil. **Antibiotics**, MDPI, v. 12, n. 2, p. 334, 2023.
- OUNIT, R. et al. Clark: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. **BMC Genomics**, BioMed Central, v. 16, n. 1, p. 236, 2015.
- PATEL, R. K.; JAIN, M. Ngs qc toolkit: A toolkit for quality control of next generation sequencing data. **PLoS One**, Public Library of Science, v. 7, n. 2, p. e30619, 2012.
- PENG, Y. et al. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. **Bioinformatics**, v. 28, n. 11, p. 1420–1428, 2012.
- PONS, J. C. et al. Dudes: a top-down taxonomic profiler for metagenomics. **Bioinformatics**, Oxford University Press, v. 35, n. 2, p. 219–227, 2019.
- PRJIBELSKI, A. et al. Using spades de novo assembler. **Current Protocols in Bioinformatics**, v. 70, p. e102, 2020.
- QIAGEN. **DNA**. 2024. Disponível em: <<https://www.qiagen.com/us/product-categories/discovery-and-translational-research/dna-rna-purification/dna-purification>>. Acesso em: 04 mar. 2025.
- QIAGEN. **DNeasy PowerSoil Pro Kits**. 2024. Disponível em: <<https://www.qiagen.com/us/products/discovery-and-translational-research/dna-rna-purification/dna-purification/microbial-dna/dneasy-powersoil-pro-kit>>. Acesso em: 04 mar. 2025.
- QUINCE, C. et al. Shotgun metagenomics, from sampling to analysis. **Nature biotechnology**, Nature Publishing Group US New York, v. 35, n. 9, p. 833–844, 2017.

- RAI, A. et al. Effect of fluorescent pseudomonas on plant growth promotion of aloe vera. **Journal of Plant Nutrition**, Taylor & Francis, v. 47, n. 7, p. 1089–1100, 2024.
- RICHARDSON, L. et al. MGnify: the microbiome sequence data analysis resource in 2023. **Nucleic Acids Research**, v. 51, n. D1, p. D753–D759, 12 2022.
- ROSSUM, G. V.; DRAKE, F. L. **Python 3 Reference Manual**. Scotts Valley, CA: CreateSpace, 2009. ISBN 1441412697.
- SALGUEIRO, V. et al. Snapshot of resistome, virulome and mobilome in aquaculture. **Science of the Total Environment**, Elsevier, v. 905, p. 166351, 2023.
- SARAO, S. K. et al. Bradyrhizobium and the soybean rhizosphere: Species level bacterial population dynamics in established soybean fields, rhizosphere and nodules. **Plant and Soil**, Springer, p. 1–16, 2024.
- SAYERS, S. et al. Victors: a web-based knowledge base of virulence factors in human and animal pathogens. **Nucleic acids research**, Oxford University Press, v. 47, n. D1, p. D693–D700, 2019.
- SCHMIEDER, R.; EDWARDS, R. **DeconSeq: A tool to remove contaminating sequences from metagenomic datasets**. 2011. Disponível em: <<https://sourceforge.net/projects/deconseq/>>. Acesso em: 04 mar. 2025.
- SCHUBERT, M.; LINDGREEN, S.; ORLANDO, L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. **BMC Research Notes**, v. 9, n. 1, p. 88, 2016.
- SEEMANN, T. Prokka: rapid prokaryotic genome annotation. **Bioinformatics**, Oxford University Press, v. 30, n. 14, p. 2068–2069, 2014.
- SEEMANN, T. **ABRicate**. 2020. Disponível em: <<https://github.com/tseemann/abricate>>. Acesso em: 04 mar. 2025.
- SENKO, O. et al. Role of humic substances in the (bio) degradation of synthetic polymers under environmental conditions. **Microorganisms**, MDPI, v. 12, n. 10, p. 2024, 2024.
- SEPPEY, M.; MANNI, M.; ZDOBNOV, E. M. Lemmi: a continuous benchmarking platform for metagenomics classifiers. **Genome Research**, Cold Spring Harbor Laboratory Press, v. 30, n. 8, p. 1208–1216, 2020.
- SOKOŁOWSKI, W. et al. In vitro screening of endophytic micromonospora strains associated with white clover for antimicrobial activity against phytopathogenic fungi and promotion of plant growth. **Agronomy**, MDPI, v. 14, n. 5, p. 1062, 2024.
- STOGIOS, P. J.; SAVCHENKO, A. Molecular mechanisms of vancomycin resistance. **Protein Science**, Wiley Online Library, v. 29, n. 3, p. 654–669, 2020.
- SZPUNAR-KROK, E. et al. Effect of nitrogen fertilization and inoculation with bradyrhizobium japonicum on nodulation and yielding of soybean. **Agronomy**, MDPI, v. 13, n. 5, p. 1341, 2023.
- TRUONG, D. T. et al. Metaphlan2: accurate profiling of microbial communities from metagenomic shotgun sequences. **Nature Methods**, Nature Publishing Group, v. 12, n. 10, p. 902–903, 2015.

- UNIPROT KNOWLEDGEBASE. **What are UniProtKB's criteria for defining a CDS as a protein?** 2024. Disponível em: <https://www.uniprot.org/help/cds_definition>. Acesso em: 04 mar. 2025.
- US DOE JOINT GENOME INSTITUTE. Initial sequencing and analysis of the human genome. **Nature**, Nature Publishing Group UK London, v. 409, n. 6822, p. 860–921, 2001.
- VAN DIJK, E. L.; JASZCZYSZYN, Y.; THERMES, C. Library preparation methods for next-generation sequencing: Tone down the bias. **Experimental Cell Research**, v. 322, n. 1, p. 12–20, 2014. ISSN 0014-4827.
- VASIMUDDIN, M. et al. Efficient architecture-aware acceleration of bwa-mem for multi-core systems. p. 314–324, 2019.
- VOLLMERS, J.; WIEGAND, S.; KASTER, A. K. Comparing and evaluating metagenome assembly tools from a microbiologist's perspective - not only size matters! **PLoS ONE**, v. 12, n. 1, p. e0169662, 2017.
- WALT, A. J. van der et al. Assembling metagenomes, one community at a time. **BMC Genomics**, v. 18, p. 521, 2017.
- WARNECKE, F. et al. Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. **Nature**, Nature Publishing Group UK London, v. 450, n. 7169, p. 560–565, 2007.
- WILHELM, R. C. et al. Ecological insights into soil health according to the genomic traits and environment-wide associations of bacteria in agricultural soils. **ISME Communications**, v. 3, n. 1, p. 1, 2023.
- WOOD, D.; LU, J.; LANGMEAD, B. Improved metagenomic analysis with kraken 2. **Genome Biology**, v. 20, p. 257, 2019.
- WOOD, D. E.; SALZBERG, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. **Genome Biology**, v. 15, n. 3, p. R46, 2014.
- XU, Z. et al. Bioinformatic approaches reveal metagenomic characterization of soil microbial community. **PLOS ONE**, Public Library of Science, v. 9, n. 4, p. 1–11, 04 2014.
- YANG, J. et al. Colonization and performance of a pyrene-degrading bacterium mycolicibacterium sp. pyr9 on root surfaces of white clover. **Chemosphere**, Elsevier, v. 263, p. 127918, 2021.
- YE, S. S. H. et al. Benchmarking metagenomics tools for taxonomic classification. **Cell**, Elsevier, v. 178, n. 4, p. 779–794, 2019.
- YIN, X. et al. Args-oap v3.0: Antibiotic-resistance gene database curation and analysis pipeline optimization. **Engineering**, v. 27, p. 234–241, 2023. ISSN 2095-8099. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S2095809922008062>>.
- ZHANG, Q. et al. The priming effect patterns linked to the dominant bacterial keystone taxa during different straw tissues incorporation into mollisols in northeast china. **Applied Soil Ecology**, Elsevier, v. 197, p. 105330, 2024.

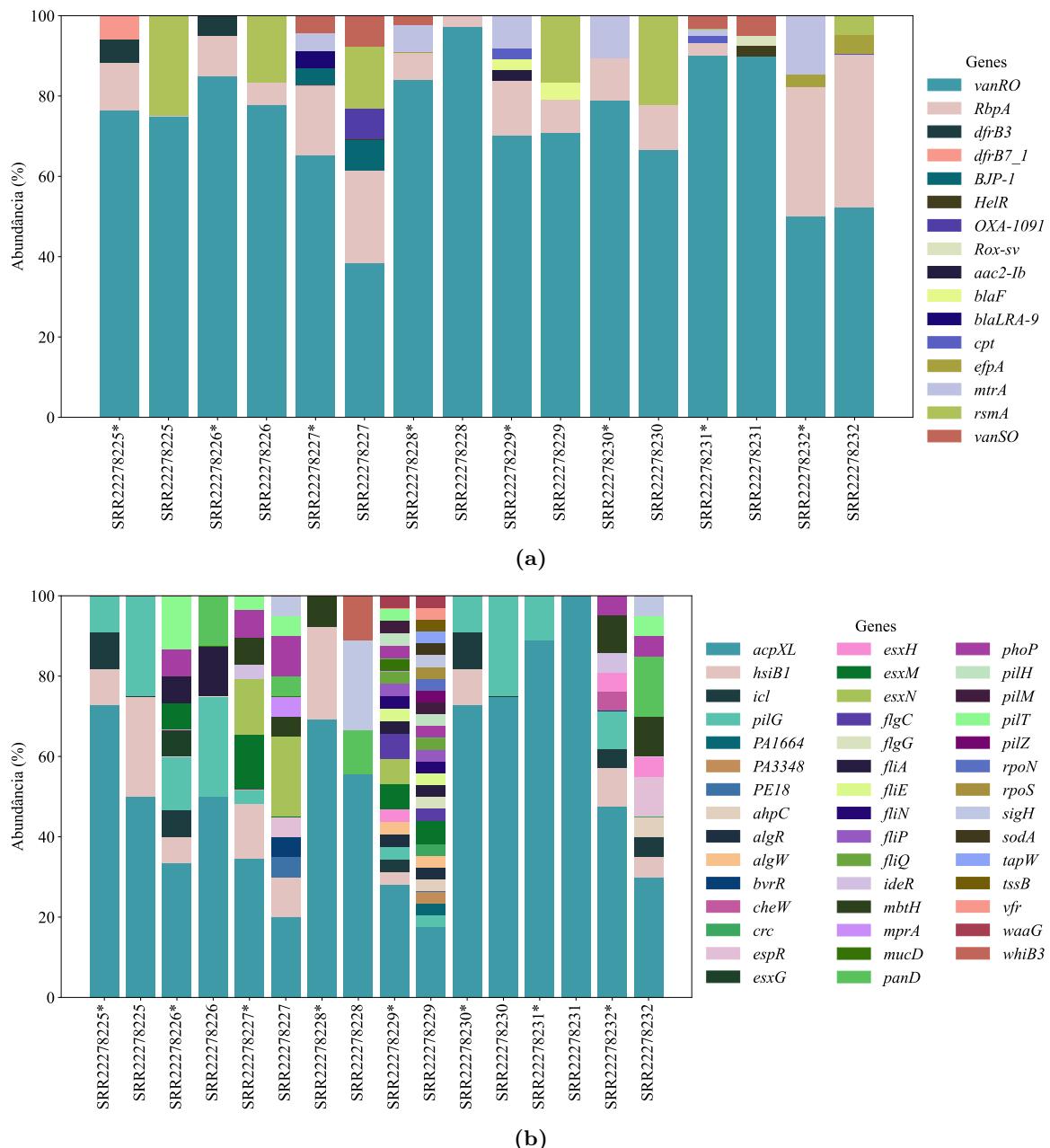
- ZHANG, S. W.; JIN, X. Y.; ZHANG, T. Gene prediction in metagenomic fragments with deep learning. **BioMed Research International**, v. 2017, p. 1–9, 2017.
- ZHANG, Y. et al. Metagenomic insights into microbial variation and carbon cycling function in crop rotation systems. **Science of The Total Environment**, Elsevier, v. 947, p. 174529, 2024.
- ZHANG, Z. et al. Benchmarking de novo assembly methods on metagenomic sequencing data. **bioRxiv**, Cold Spring Harbor Laboratory, p. 2022–05, 2022.
- ZHOU, C. et al. Mvirdb—a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. **Nucleic acids research**, Oxford University Press, v. 35, n. suppl_1, p. D391–D394, 2007.
- ZHU, W. **Improvement of ab initio methods of gene prediction in genomic and metagenomic sequences**. [S.l.]: Georgia Institute of Technology, 2010.

APÊNDICE – COMPARAÇÕES COM ORDINE ET AL. (2023)

A Figura 10 compara os genes aqui identificados com aqueles encontrados originalmente por Ordine et al. (2023), com respeito tanto aos ARGs [Figura 10(a)] quanto aos VFs [Figura 10(b)]. A partir da Figura 10(a), observa-se uma diferença significativa dentre os genes identificados aqui em comparação com (ORDINE et al., 2023), sendo a ampla presença de *vanRO* o principal padrão constatado. Vale destacar que alguns genes foram encontrados apenas aqui (e.g., o *rsmA*), enquanto outros genes (e.g., o *mtrA*) foram identificados apenas em Ordine et al. (2023). Por sua vez, a partir da Figura 10(b), verifica-se que a variabilidade dos genes identificados é ainda maior que a observada através dos ARGs, sendo a ampla presença de *acpXL* o principal padrão constatado. Todavia, apesar dos genes identificados aqui diferirem em abundância e em diversidade daqueles encontrados por Ordine et al. (2023), a *pipeline* implementada fez uso de ferramentas atualizadas e seguiu padrões já reconhecidos. Dessa maneira, pode-se argumentar que tais diferenças podem ter ocorrido pelos seguintes motivos:

- **Versões diferentes:** As versões de alguns *softwares* empregados não foram devidamente detalhadas por Ordine et al. (2023), assim como as versões dos bancos de dados CARD e VFDB utilizadas pelo ABRicate.
- **Parâmetros distintos:** A parametrização considerada nos diferentes *softwares* utilizados por Ordine et al. (2023) não foi devidamente apresentada.
- **Erros de implementação:** Os *scripts* e/ou códigos não foram disponibilizados publicamente por Ordine et al. (2023), impossibilitando assim a detecção de quaisquer erros cometidos.

Figura 10 – Comparação entre os genes encontrados aqui e aqueles de Ordine et al. (2023) (identificados pelo símbolo ‘*’). (a) Genes de resistência a antibióticos (ARGs). (b) Genes de fatores de virulência (VFs).



Fonte: Autoria própria.