DESENVOLVIMENTO DE BANCOS DE DADOS PARA ANÁLISE METAGENÔMICA DE MICROBIOMAS DE CÃES: UMA ABORDAGEM COM KRAKEN2 E BOWTIE2

PAULO SALLAROLA TAKAO; PAMELA SOUZA CORRÊA; JULIO FRANZ MOURA; DAVID ACIOLE BARBOSA; FABIANO BEZERRA MENEGIDIO

RESUMO

A análise metagenômica desempenha um papel crucial na compreensão das comunidades microbianas e seu impacto ambiental, desfrutando de uma relevância especial na medicina veterinária para diagnóstico, tratamento e prevenção de doenças em animais de estimação. Este estudo teve como objetivo principal criar bancos de dados integrados para as ferramentas Bowtie2 e Kraken2, consolidando os genomas do cão e do humano em um recurso atualizado e de fácil acesso. Com o avanço tecnológico, a análise metagenômica do microbioma emerge como uma ferramenta promissora na rotina veterinária, especialmente considerando a alta prevalência de cães nas clínicas veterinárias, sabendo-se que apenas no Brasil temos em torno de 58,1 milhos de cães, sendo o segundo animal de estimação mais presente nos lares, perdendo apenas para as aves. No entanto, as etapas bioinformáticas necessárias, como a remoção do genoma do hospedeiro e de contaminantes humanos, demandam tempo e recursos computacionais consideráveis. Para superar esse desafio, foram desenvolvidos bancos de dados específicos para cada ferramenta, reduzindo significativamente o tempo de análise e assegurando a atualização contínua dos genomas. Enquanto o Bowtie2 realiza o alinhamento exato de sequências, o Kraken2 utiliza sequências menores (k-mers) para uma classificação taxonômica mais rápida e eficiente. Após a construção dos bancos de dados, foram conduzidos testes em uma biblioteca de metagenoma canino, evidenciando uma alta porcentagem de alinhamento e uma remoção eficaz de leituras pertinentes ao cão ou humano. Apesar dos desafios enfrentados devido à falta de documentação detalhada, os bancos de dados criados se mostraram viáveis e reproduzíveis, oferecendo uma contribuição significativa para estudos futuros em análise metagenômica veterinária.

Palavras-chave: Canis lupus familiaris, Metagenoma, Kraken2-build, Bowtie2-build, Contaminantes.

1 INTRODUÇÃO

A análise metagenômica é o processo utilizado para classificar o metagenoma (MARCHESI; RAVEL, 2015), sendo o metagenoma uma derivação da genômica que é o estudo dos genes de um organismo, então na tradução literal metagenômica seria “além do genoma”, ou seja, na metagenômica não analisamos apenas um genoma e sim todo os genomas contidos em uma amostra (GILBERT; DUPONT, 2011), podendo esses genomas serem de microrganismos e até fragmentos de DNA livres no ambiente (amostra), genes de elementos estruturais microbianos, vírus, fagos, toxinas e tudo mais que possua material genético (BERG, *et al.*, 2020; HANDELSMAN *et al.*, 1998; MERRIAM-WEBSTER, 2023; WHIPPS; LEWIS; COOKE, 1988), então podemos descrever a análise metagenômica como a análise de uma coleção de genomas de microrganismos e suas condições ambientais em um determinado

DOI: 10.51189/conbiotec2024/33826

ambiente por uma técnica de sequenciamento genético independente da utilização de cultura para crescimento e análise (HANDELSMAN *et al.*, 1998; MARCHESI; RAVEL, 2015; MERRIAM-WEBSTER, 2023; WHIPPS; LEWIS; COOKE, 1988). Portanto com a tecnologia de análise metagenômica podemos descobrir um mundo totalmente novo, já que se sabe que apenas pouco mais de 1% de todos os microrganismos existentes são passiveis de serem cultivados (COWAN, 2000; HANDELSMAN, 2004; MERRIAM-WEBSTER, 2023). Uma das principais aplicações da análise metagenômica é o estudo do microbioma que foi definido pela primeira vês por Whipps at al. (1988), como sendo uma comunidade de microrganismos (WHIPPS; LEWIS; COOKE, 1988), mas essa definição foi se alterando ao longo dos anos onde chegou-se a definição de microbioma como todo o material genético (DNA) da microbiota de um determinado ambiente e sua área de atividade, incluindo elementos estruturais interno e externos dos microrganismo ou seja é quando se analisa não apenas a comunidade de micróbios que colonizam este ambiente, mas também a expressão dos genes desta microbiota e como eles influenciam aquele ambiente (BERG *et al.*, 2020; MERRIAM-WEBSTER, 2023).

Neste contexto, a análise metagenômica do microbioma emerge como uma ferramenta promissora na rotina veterinária, especialmente considerando a alta prevalência de cães (*Canis lúpus familiaris*) nas clínicas veterinárias e de acordo com o relatório de 2023 da AMBIPET (Associação brasileira da indústria de produtos para animais de estimação) o Brasil é o segundo maior pais em população de cães no mundo, ficando atrás apenas do Estados Unidos, tendo aproximadamente de 58,1 milhões e 900 milhões de cães respectivamente (“Mercado Pet Brasil 2023”, 2023).

Com o avanço das tecnologias de sequenciamento genético e da diminuição dos custos para a sua execução (BHARTI; GRIMM, 2021), a análise metagenômica do microbioma está cada dia mais perto da utilização na rotina da clínica veterinária de pequenos animais sendo uma promissora ferramenta de diagnostico, tratamento e prevenção de doenças, sendo que na medicina humana já está bastante difundida principalmente para a análise do microbioma de intestino, boca e pele (THE INTEGRATIVE HMP (IHMP) RESEARCH NETWORK CONSORTIUM, 2019), como o cão é o animal com a maior presença na rotina do médico veterinário de pequenos animais o conhecimento do seu microbioma é de fundamental importância (“Mercado Pet Brasil 2023”, 2023; “Número de cães e gatos no Brasil deve chegar a mais de 100 milhões em 10 anos”,). Porem para a realização de uma análise metagenômica várias etapas de bioinformática são necessários, sendo uma das principais etapas a retirada do genoma do hospedeiro, sendo nesse caso o cão (*Canis lupus familiaris*) e de genomas contaminantes do ser humano (*Homo sapiens*), assim conseguimos diminuir consideravelmente o volume de dados a serem processados, já que se sabe que em uma amostra de sequenciamento metagenômico a quantidade de material genético do hospedeiro e do contaminante pode ser mais de 99% da amostra.

Para a retirada do genoma do hospedeiro e de contaminantes a ferramenta de bioinformática a ser utilizada possui um banco de dados (BD) previamente indexado específico para cada organismo que muitas vezes é disponibilizado gratuitamente pela ferramenta, no entanto esses BD são disponibilizados de forma individual, no nosso caso um BD para cão e um BD para ser humano fazendo com que a ferramenta tenha que ser utilizada duas vezes tendo uma perda de tempo desnecessária, já que algumas ferramentas podem demorar muitas horas para realizar apena uma análise fora isso muitas vezes esses banco de dados já estão defasados sem atualização do genoma. Duas das ferramentas de uso corriqueiro para essa função, são o Bowtie2 e Kraken2 (LANGMEAD; SALZBERG, 2012; LU *et al.*, 2022; WOOD; SALZBERG, 2014).

Bowtie2, realiza o alinhamento de sequencias exatas de acordo com o seu BD de referência, esse alinhamento ocorre em 4 etapas onde na etapa 1 ele irá coletar sementes (inicio) de sequencias a serem analisadas, na etapa 2 ele irá alinhara as sementes são extraídas da

DOI: 10.51189/conbiotec2024/33826

amostra e alinhadas de forma não lacunar, na etapa 3 as sementes são posicionadas no genoma de referência e na etapa 4 essas sementes são estendidas para um alinhamento completo (LANGMEAD; SALZBERG, 2012; LU *et al.*, 2022; RUMBAVICIUS; ROUNGE; ROGNES, 2023). As sementes não alinhadas e suas respectivas sequencias são arquivadas separadamente para posterior análise do microbioma (LANGMEAD; SALZBERG, 2012). Devido ser dividido em 4 etapas torna a ferramenta mais lenta porem com um baixo consumo de memória ram (RUMBAVICIUS; ROUNGE; ROGNES, 2023). Essa ferramenta possui os bancos de dados já indexados do cão e do Homem disponíveis publicamente, porem com genomas defasados e de forma individual, sendo um genoma por banco de dados, acabando que temos que executar duas vezes a ferramenta, primeiro com um banco de dados e depois com o outro (LANGMEAD *et al.*, 2023).

Para criação do banco de dados Bowtie2, é utilizado o bowtie2-build que constrói um índice Bowtie a partir de sequencias de DNA contidos em um arquivo do tipo ".fasta", produzindo um conjunto de 6 arquivos com o sufixo ".bt2"(LANGMEAD; SALZBERG, 2012). Kraken2 é uma ferramenta de classificação taxonômica muito utilizada na metagenômica que diferente do Bowtie2 que realiza o alinhamento completo com uma sequência de DNA, ela utiliza sequencias menores com tamanho k (k-mers) sendo o tamanho padrão pré-definido de "k" igual a 50 nucleotídeos, que pode ser alterado de acordo com a necessidade, assim deixando a ferramenta muito mais ágil podendo analisar até 1,5 milhões de leituras (reads) por minuto com muita precisão (LANGMEAD *et al.*, 2023; LANGMEAD; SALZBERG, 2012; LU *et al.*, 2022; RUMBAVICIUS; ROUNGE; ROGNES, 2023; WOOD; SALZBERG, 2014). Essa ferramenta até a pouco tempo não era utilizada para a retirada de hospedeiro e contaminantes de amostras de metagenomas, somente após a sua última atualização foi implementado essa funcionalidade e vem demonstrando uma ótima opção para isso, sendo mais rápida e com uma boa sensibilidade após testes realizados por nossa equipe. Diferente da ferramenta anterior, essa não possui bancos de dados com apenas um ou com os dois genomas juntos, por ser uma ferramenta utilizada corriqueiramente apenas para análise taxonômica de metagenomas, ela tem apenas BD com genomas de múltiplos organismos. Para a criação de BD Kraken2 é utilizado o kraken2-build que criará um diretório com 3 arquivos com o sufixo ".k2d", que é criado a partir de arquivos .fasta contendo os genomas de referência.

O Objetivo desse trabalho foi criar bancos de dados específicos para cada ferramenta sendo elas Bowtie2 e Kraken2 que são disponibilizadas gratuitamente por seus criadores, com o genoma do hospedeiro e contaminante em um único BD com os genomas atualizados para as duas espécies *Canis lupus familiaris* e *Homo sapiens*.

2 MATERIAIS E MÉTODOS

Esse projeto faz parte de um estudo maior para a uma dissertação de mestrado.

Para a confecção do banco de dados foi utilizado um notebook Acer Nitro 5 com processador intel core I5 11ª geração com 6 núcleos e 12 threads, 42 Gb de memória ram, placa de vídeo GeForce GTX 1650 e sistema operacional Linux Mint 21.2 Cinnamon, baseado em Ubuntu.

Para que as ferramentas pudessem ser executadas, foi instalado um pacote de de softwares de pesquisa biomédicas o Bioconda através da linha de comando utilizando o comando e em seguida foram instaladas as ferramentas Bowtie2, Bowtie-build, Kraken2 e Kraken2_build Iniciamos a confecção dos BD, o primeiro passo foi realizado o download dos arquivos dos genomas de referencia atualizados no formato fasta (.fna) do cão e do Homem, disponíveis publicamente no banco de dados do *RefSeq* (Banco de Dados de Sequência de Referência) do NCBI (National Center for Biotechnology Information) (tabela 1) (BETHESDA (MD): NATIONAL LIBRARY OF MEDICINE (US), 1988; "RefSeq: NCBI Reference Sequence

DOI: 10.51189/conbiotec2024/33826

Database", 2023).

Tabela 1: Dados dos genomas utilizados para a confecção dos bancos de dados de cão e ser humano, extraídos do banco de dados publico do RefSeq NCBI (https://www.ncbi.nlm.nih.gov/refseq/) (BETHESDA (MD): NATIONAL LIBRARY OF MEDICINE (US), 1988; "RefSeq: NCBI Reference Sequence Database", 2023).

Nome Científico	Biblioteca	RefSeq	Modifica dos	Tamanho (Mb)	Nível	Data Lançamento
Canis lupus familiaris	Dog10k_Boxer_ Tasha	GCF_000002285.5	Boxer	2341	Cromossomo	Outubro 2020
Homo Sapiens	GRCh38.p14	GCF_000001405.40		3099	Cromossomo	Fevereiro 2022

Após o download dos arquivos, iniciamos a confecção dos BD pelo Bowtie2 pela linha de comando, no ambiente conda, utilizando os parâmetros de 12 threads (-- threads 12) forçando a utilização da capacidade máxima do computador para que fosse executado com o menor tempo possível e -f (local do arquivo fasta com o genoma) separando os arquivos com os respectivos genomas com uma virgula (,) e em seguida separado por um espaço o local de destino do novo banco de dados, comando completo a seguir:

```
$ bowtie2-build --threads 12 -f \
pasta_origem/genoma_cao.fna,pasta_origem/genoma_humano.fna \
pasta_destino_BD/BD_cao_humano_bt2/bt2
```

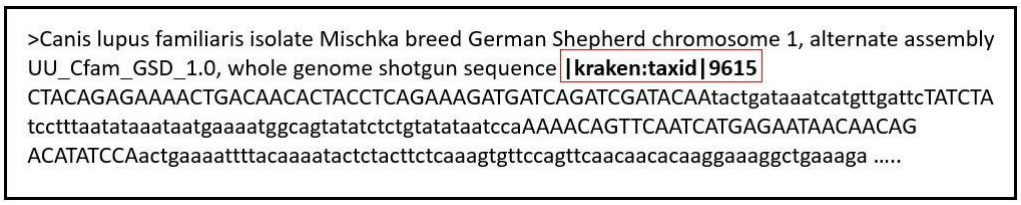
Para a confecção do banco de dado Kraken2, tivemos que realizar uma alteração no arquivo fasta para que o programa kraken2-build consiga identificar corretamente cada sequencia de acordo com o sua taxonomia, então colocamos o numero de identificação da taxonomia da espécie (taxid) na linha de identificação de cada sequencia contida no arquivo (fig 1), o numero de identificação taxonomica do *Canis lupus familiaris* (taxid: 9615) e do *Homo sapiens* (taxid: 9606) foram prospectors da base de dados pública de taxonomia do NCBI (BETHESDA (MD): NATIONAL LIBRARY OF MEDICINE (US), 1988; "Taxonomy Database",) e como cada arquivo pode ter milhares de linhas de identificação, utilizamos o software SeqFu na linha de comando, que é utilizado para manipulação de arquivos fasta, foi acrescentado a linha de identificação o seguinte texto "|kraken:taxid|000000" (figura 2), para isso utilizamos o seguinte comando:

```
$ seqfu cat --append "|kraken:taxid|(numero_da_taxonomia)" ~/local_arquivo fasta/arquivo_fna \ > ~/destino_arquivo_editado/arquivo_taxid.fna
```

Figura 1: Primeira linha do arquivo fasta do genoma de *Canis lupus familiaris* onde se destaca a linha de identificação onde deve conter a identificação da taxonomia para ferramenta kraken2-build.

```
>Canis lupus familiaris isolate Mischka breed German Shepherd chromosome 1, alternate assembly
UU_Cfam_GSD_1.0, whole genome shotgun sequence
CTACAGAGAAAACTGACAACACTACCTCAGAAAGATGATCAGATCGATACAAtactgataaatcatgttgattcTATCTA
tcctttaatataaataatgaaaatggcagtatatctctgtatataatccaAAAACAGTTCAATCATGAGAATAACAACAG
ACATATCCAactgaaaattttacaaaatactctacttctcaaagtgttccagttcaacaacacaaggaaaggctgaaaga .....
```

Figura 2: primeira linha do arquivo fasta do genoma de *Canis lupus familiaris* com o Taxid: 9615 em seu devido local para identificação pela ferramenta kraken2-build.



Com os arquivos fastas já preparados com os respectivos taxid iniciamos a construção da biblioteca para Kraken2 com os 2 genomas. A criação desse banco de dados consiste em 4 etapas, onde iremos criar uma biblioteca com os dois bancos de dados, em seguida será realizado o download da arvore filogenética para os dois genomas com a identificação de todas as suas taxonomias em seguida é realizada a compilação do banco de dados e por ultimo realizamos a limpeza dos arquivos desnecessários criados durante o processo.

Etapa 1: Criação de uma biblioteca com os dois genomas, deve ser realizado os comandos 2x um para cada genoma e arquivá-los na mesma pasta de saída:
$ kraken2-build --add-to-library ~/local_arquivo_fasta/cao_taxid.fasta --db ~/cao_humanodb --threads 12 $ kraken2-build --add-to-library ~/local_arquivo_fasta/humano_taxid.fasta –db ~/cao_humanodb --threads 12

Etapa 2: Download da taxonomia: $ kraken2-build --download-taxonomy --db ~/cao_humano

Etapa 3: Compilação do banco de dados final para Kraken2: $ kraken2-build --build --db ~/cao_humanodb

Etapa 4: Realização da limpeza de arquivos desnecessários utilizados para a criação do banco de dados: $ kraken2-build --clean --db ~/cao_humanodb

3 RESULTADOS E DISCUSSÃO

Após a compilação dos bancos de dados para ambas as ferramentas nos realizamos a inspeção "*inspect"* das mesmas onde foi possível identificar que os bancos de dados estavam completos e contendo os dois genomas e aptos a serem utilizados para a retirada de hospedeiro e contaminantes de bibliotecas de metagenoma e como teste complementar os dois bancos de dados foram testadas em uma biblioteca de metagenoma de cão com o SRA ID SRR19324877 disponíveis publicamente no *sequence reads arquive* (SRA) do NCBI.

Tabela 2: Teste realizado em uma biblioteca de metagenoma (SRR19324877) que possui 250.223.105 leituras, onde foi possível observar que os dois bancos de dados executados nas ferramentas Bowtie2 e Kraken2 com a configuração padrão das ferramentas, foi capas de ter uma boa porcentagem de alinhamentos e deixando apenas leituras de genomas não identificados como cão e ser humano.

Ferramenta	Leituras Brutas	Leituras alinhadas após análise	% de alinhamento	Leituras para análise do microbioma
Bowtie2	250223105	234749227	93,81%	15473878
Kraken2	250223105	250014348	99,91%	208757

A maior dificuldade para a construção dos bancos de dados foi a falta de artigos

DOI: 10.51189/conbiotec2024/33826

disponíveis sobre o tema e os manuais das ferramentas são muito sucintos nas explicações, faltando alguns dados fundamentais como a colocação da taxid no arquivo fasta para a criação do BD Kraken-2, a maioria das informações que tivemos foram a partir de fóruns e blogs do gênero, mas mesmo assim com alguns dados não tão precisos, a criação desses bancos de dados só foi possível após muitos testes de diferentes configurações para ambas as ferramentas.

4 CONCLUSÃO

Concluímos que os bancos de dados criados são viáveis e demonstraram êxito na realização do que foi proposto, sendo reprodutíveis para as mais diversas finalidades e genomas variados.

REFERÊNCIAS

BERG, G.; RYBAKOVA, D.; FISCHER, D.; CERNAVA, T.; VERGÈS, M.-C. C.; CHARLES, T.; CHEN, X.; COCOLIN, L.; EVERSOLE, K.; CORRAL, G. H.; KAZOU, M.; KINKEL, L.; LANGE, L.; LIMA, N.; LOY, A.; MACKLIN, J. A.; MAGUIN, E.; MAUCHLINE, T.; MCCLURE, R.; MITTER, B.; RYAN, M.; SARAND, I.; SMIDT, H.; SCHELKLE, B.; ROUME, H.; KIRAN, G. S.; SELVIN, J.; SOUZA, R. S. C. de; VAN OVERBEEK, L.; SINGH, B. K.; WAGNER, M.; WALSH, A.; SESSITSCH, A.; SCHLOTER, M. Microbiome definition re-visited: old concepts and new challenges. **Microbiome**, v. 8, n. 1, p. 103, 30 jun. 2020.

BETHESDA (MD): NATIONAL LIBRARY OF MEDICINE (US). **National Center for Biotechnology Information (NCBI)**. Disponível em: <https://www.ncbi.nlm.nih.gov/>. Acesso em: 2 nov. 2023.

BHARTI, R.; GRIMM, D. G. Current Challenges and Best-Practice Protocols for Microbiome Analysis. **Briefings in Bioinformatics**, v. 22, n. 1, p. 178–193, 18 jan. 2021.

COWAN, D. A. Microbial Genomes – the Untapped Resource. **Trends in Biotechnology**, v. 18, n. 1, p. 14–16, jan. 2000.

GILBERT, J. A.; DUPONT, C. L. Microbial Metagenomics: Beyond the Genome. **Annual Review of Marine Science**, v. 3, n. 1, p. 347–371, 15 jan. 2011.

HANDELSMAN, J. Metagenomics: Application of Genomics to Uncultured Microorganisms. **Microbiology and Molecular Biology Reviews**, v. 68, n. 4, p. 669–685, dez. 2004.

HANDELSMAN, J.; RONDON, M. R.; BRADY, S. F.; CLARDY, J.; GOODMAN, R. M. Molecular Biological Access to the Chemistry of Unknown Soil Microbes: A New Frontier for Natural Products. **Chemistry & Biology**, v. 5, n. 10, p. R245-249, out. 1998.

LANGMEAD, B.; KIN, D.; CHARLES, R.; CHEN, N.-C.; WILKS, C.; ANTONESCU, Va. **Bowtie 2 Fast and sensitive read alignmentBowtie 2**13 out. 2023. Disponível em: <https://bowtie-bio.sourceforge.net/bowtie2/manual.shtml#the-bowtie2-build-indexer>. Acesso em: 5 nov. 2023.

LANGMEAD, B.; SALZBERG, S. L. Fast Gapped-Read Alignment with Bowtie 2. **Nature Methods**, v. 9, n. 4, p. 357–359, abr. 2012.

DOI: 10.51189/conbiotec2024/33826

LU, J.; RINCON, N.; WOOD, D. E.; BREITWIESER, F. P.; POCKRANDT, C.; LANGMEAD, B.; SALZBERG, S. L.; STEINEGGER, M. Metagenome Analysis Using the Kraken Software Suite. **Nature Protocols**, v. 17, n. 12, p. 2815–2839, dez. 2022.

MARCHESI, J. R.; RAVEL, J. The vocabulary of microbiome research: a proposal. **Microbiome**, v. 3, n. 1, p. 31, 30 jul. 2015.

Mercado Pet Brasil 2023. **AMBIPET associação brasileira da industria de produtos para naimais de estimação**2023. Disponível em: <https://abinpet.org.br/dados-de-mercado/>. Acesso em: 2 nov. 2023.

MERRIAM-WEBSTER. **Definition of MICROBIOME**. Disponível em: <https://www.merriam-webster.com/dictionary/microbiome>. Acesso em: 17 nov. 2023.

Número de cães e gatos no Brasil deve chegar a mais de 100 milhões em 10 anos. **SIDAN Saude Animal**[s.d.]Disponível em: <https://sindan.org.br/release/numero-de-caes-e-gatos-no-brasil-deve-chegar-a-mais-de-100-milhoes-em-10-anos/#:~:text=A%20popula%C3%A7%C3%A3o%20total%20de%20c%C3%A3es,milh%C3%B5es%20de%20animais%20at%C3%A9%202030.>. Acesso em: 2 nov. 2023.

RefSeq: NCBI Reference Sequence Database. **National Library of Medicine**11 set. 2023. Disponível em: <https://www.ncbi.nlm.nih.gov/refseq/>. Acesso em: 5 nov. 2023.

RUMBAVICIUS, I.; ROUNGE, T. B.; ROGNES, T. HoCoRT: Host Contamination Removal Tool. **BMC Bioinformatics**, v. 24, n. 1, p. 371, 2 out. 2023.

Taxonomy Database. **National Library of Medicine**[s.d.]Disponível em: <https://www.ncbi.nlm.nih.gov/taxonomy>. Acesso em: 5 nov. 2023.

THE INTEGRATIVE HMP (IHMP) RESEARCH NETWORK CONSORTIUM. The Integrative Human Microbiome Project. **Nature**, v. 569, n. 7758, p. 641–648, maio 2019.

WHIPPS, J.; LEWIS, K.; COOKE, R. Mycoparasitism and plant disease control. **Fungi Biol Control Syst.**, p. 161–187, 1988.

WOOD, D. E.; SALZBERG, S. L. Kraken: Ultrafast Metagenomic Sequence Classification Using Exact Alignments. **Genome Biology**, v. 15, n. 3, p. R46, 3 mar. 2014.

DOI: 10.51189/conbiotec2024/33826