Análises computacionais para identificação de sequências virais em dados metagenômicos

Palavras-Chave: MICROBIOTA, DEEP LEARNING, SEQUENCIAMENTO DE PRÓXIMA GERAÇÃO.

Autores(as):

Monyque Karoline de Paula Silva, Ilum – Escola de Ciência.

Profº. Drº. Leandro Nascimento Lemos (orientador), Ilum – Escola de Ciência.

INTRODUÇÃO:

Os vírus são as entidades biológicas mais diversas e abundantes, com uma população estimada em 10^{39} , a qual habita diversos ambientes e infecta espécies de todos os domínios da vida por meio do reconhecimento e ligação de receptores específicos da célula hospedeira para infiltração e exploração da maquinaria molecular do hospedeiro. Essencialmente, o viroma intestinal é composto por vírus eucarióticos capazes de se replicar nas células humanas, assim como por vírus procariontes como os bacteriófagos que se infectam em bactérias intestinais (Lecuit; Eloit, 2017; Garmaeva *et al*., 2019).

Numa comunidade microbiana, os bacteriófagos (conhecidos como os fagos) são os principais componentes do material genético viral, sendo os fagos 10 vezes mais presentes do que as bactérias. Eles são vírus que podem infectar bactérias para sua replicação viral, mas em algumas situações, podem beneficiar as populações de bactérias no hospedeiro, demonstrando o seu impacto crucial na composição da comunidade microbiana. Eles medeiam a transferência lateral de genes e alteram o metabolismo do hospedeiro por meio de genes metabólicos auxiliares (AMGs) (Wommack *et al.*, 2000; Al-Shayeb *et al.*, 2020; Wu *et al.,* 2021). Compreender os fagos e as AMGs significa entender melhor os papeis ecológicos dos vírus e seus mecanismos de funcionamento.

Há apenas 2640 genomas de fagos completamente sequenciados, demonstrando que a maioria dos fagos tem sua composição desconhecida, devido à dificuldade de cultivo isolado do vírus. Uma abordagem para explorar estes vírus é por meio do sequenciamento de nova geração (NGS), a qual realiza a leitura de cada base (A, C, T, G) individual do DNA, em larga escala, gerando *reads* com abundância de dados genéticos dependendo apenas do material genômico viral extraído diretamente de uma amostra (Al-Shayeb *et al.*, 2020).

Estes sequenciamentos possibilitam a leitura de pares de base curtos, *short-reads*, e pares de base longos, *long-reads*. As *short-reads*, são leituras de pequenos fragmentos, 50 à 600 pb, por meio da quebra do DNA em pequenos fragmentos, anexando adaptadores em cada extremidade destes para ser possível a leitura completa do DNA nos instrumentos de sequenciamento, como NovaSeq, HiSeq, NextSeq e MiSeq da Illumina (Goodwin *et al*., 2016; Jeon *et al.*, 2019). Elas são leituras econômicas e

eficientes com diversos *pipelines,* usualmente, aplicado para a contagem da abundância de sequências específicas, a identificação de variantes em sequências bem conservadas ou para traçar o perfil da expressão de transcrições específicas (Amarasinghe, 2020).

Entretanto, este tipo de leitura pode dificultar a reconstrução e contagem das sequências originais do metagenoma dado que os polímeros naturais de ácidos nucleicos abrangem oito ordens de grandeza de comprimento. Desse modo, tecnologias como o da Pacific Biosenses (PacBio) foram desenvolvidas para a realização de leituras superiores a 10 kb, sendo utilizadas para a leitura de trechos contíguos e lidar com regiões complexas do genoma para a montagem de novo genoma e detecção de variação estrutural (Amarasinghe, 2020).

O uso destas técnicas para o sequenciamento de DNA possibilita a identificação com alta qualidade das regiões complexas e contíguos dos fagos. Entretanto, não se é estabelecido um *pipeline computacional* que utilize os dois dados de sequenciamento como *input* para a identificação dos vírus, o que dificulta no tratamento destes dados de maneira eficiente e com alta acurácia.

Logo, este projeto propõe um pipeline computacional para o desenvolvimento de uma ferramenta de automação para identificação e dereplicação de bacteriófagos reconstruídos a partir de genomas virais montados em metagenoma (vMAGs) usando dados dos sequenciadores Illumina NovaSeq e PacBio SequelIIE. Genomas de bacteriófagos serão identificados usando ferramentas de aprendizado profundo voltadas para caracterização de novas espécies e construção de uma interface amigável para iniciantes em bioinformática. Espera-se que o projeto facilite o processo de caracterização de bacteriófagos para os grupos de pesquisa em bioinformática do Laboratório Nacional de Biorrenováveis (LNBR) do Centro Brasileiro de Pesquisa em Energia e Materiais (CNPEM), bem como contribua para a formação de novos recursos humanos no campo emergente de Biologia Computacional e Virologia Metagenômica na Ilum - Faculdade de Ciências, CNPEM.

METODOLOGIA:

Para o desenvolvimento do projeto estão sendo utilizados metagenomas de espécies animais nativas de biomas brasileiros a partir do sequenciamento em larga. Na Figura 1, há o esquemático do *pipeline* que consiste em três grandes etapas: o processamento de dados das leituras de *short-reads*, o processamento das *long-reads* e a integração dos dados das diferentes tecnologias.

Todo as etapas estão sendo organizadas no sistema de fluxo de automatização de dados Nextflow, o qual tem seu pipeline construído de acordo com processos individuais que são configurados com requisitos de entradas e declarações de saída. A execução de um processo começa quando todos os seus requisitos de entrada são atendidos. Ao especificar a saída de um processo como a entrada de outra etapa, uma conexão lógica e sequencial entre os processos é criada (Tommaso, 2017).

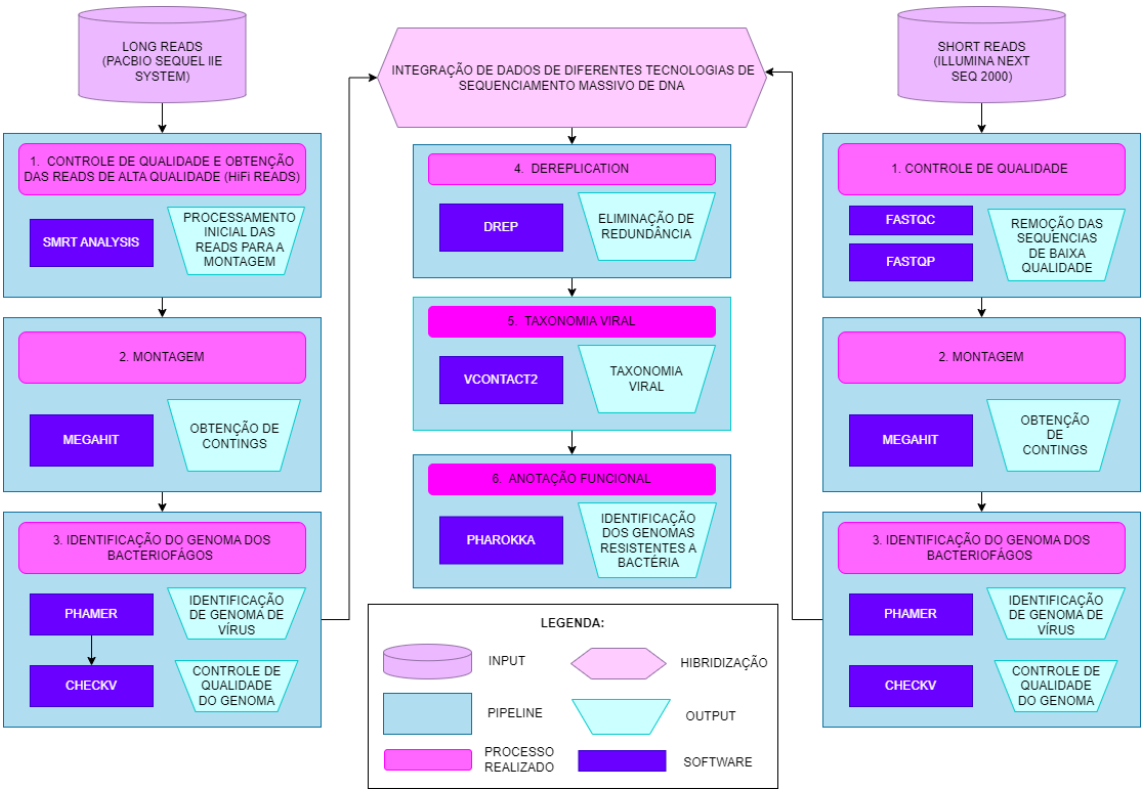

Figura 1 – Pipeline para a exploração de fagos na microbiota intestinal de animais do bioma brasileiro por meio de sequenciamento de short e long-reads. O esquema ilustra o controle de qualidade (1), a montagem dos genomas (2), a identificação dos genomas dos bacteriofágos (3), eliminação de redundância genômica (dereplicação) (4), identificação taxonômica (5) e anotação funcional (6). Imagem de autoria própria.

O pré-processamento de *short-reads* do metagenoma está ocorrendo por meio do FASTQC para checagem da qualidade da leitura. A remoção de adaptadores e sequencias de baixa qualidade, ocorreu pelo software FASTP (Chen, 2018).

As *reads de* alta qualidade das *short-reads* dos metagenomas poderão ser montados a partir de estratégias *de novo* usando a ferramenta megahit v.1.2.0 (Li, 2015) com parâmetro de tamanho mínimo de 1.000 bp e tamanho k-mer de 21 a 141 com incremento de 20.

A identificação dos bacteriófagos ocorrerá por meio do software PhaMer (Shang, 2022), o qual utiliza um modelo de análise de contextos advindo da Linguagem de Processamento Natural (NPL) para aprender padrões associados a proteínas em fagos. De modo a inferir a qualidade e a integridade da análise, utiliza-se o sofftware CheckV que observa os genomas virais fechados, estimando a integridade dos fragmentos do genoma e removendo as regiões hospedeiras flanqueadoras de provírus integrados (Nayfach, 2021).

Após o tratamento isolado de cada técnica, os dados serão postos para a dereplicação para compreender a redundância de genomas que foram identificados em ambas as técnicas por sequências por meio do software dRep (Olm, 2017). Estabelecido este processo, realiza-se a taxonomia viral dos bacteriófagos através do software vContac2 (Jang, 2019), um algoritmo de classificação taxonômica para determinar a taxonomia do genoma dos bacteriófagos.

Por fim, as análises se direcionam para a determinação da anotação funcional dos genomas dos fagos por meio do software Pharokka (Boura, 2023), uma ferramenta exclusiva para a análise dos bacteriófagos que permitem a utilização de features como o pequeno tamanho do gene, alta densidade de codificação e códons de início alternativos, ajudando a identificar as AMGs para a classificação destes genes.

RESULTADOS E DISCUSSÃO:

O *pipeline computacional* é funcional, todos os softwares atuam como bons identificadores e classificadores de fagos. Atualmente, o projeto está alocando por meio do *workflow* do *Nextflow* uma interface automatizada, sendo decidido a separação do processo em duas etapas: (1) o pré-tratamento de dados que consiste no controle de qualidade e montagem dos genomas e a (2) identificação, *dereplicação* e caracterização dos fagos.

BIBLIOGRAFIA

LECUIT, Marc; et al. **The Viruses of the Gut Microbiota**. In: Microbiota in Gastrointestinal pathophysiology. 2017. p. 179-183. DOI: 10.1016/B978-0-12-804024-9.00021-5. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7173437/. Acesso em: 6 jan. 2017.

GARMAEVA, S.; SINHA, T.; KURILSHIKOV, A.; et al. **Studying the gut virome in the metagenomic era: challenges and perspectives.** BMC Biol, v. 17, n. 1, p. 84, 2019. DOI: 10.1186/s12915-019-0704-y. Disponível em: https://doi.org/10.1186/s12915-019-0704-y. Acesso em: 7 ago. 2024.

AL-SHAYEB, Basem; SACHDEVA, Rohan; CHEN, Lin-Xing; et al. **Clades of huge phages from across Earth's ecosystems**. Nature, v. 578, p. 425-431, 2020. DOI: 10.1038/s41586-020-2007-4. Disponível em: https://doi-org.ez106.periodicos.capes.gov.br/10.1038/s41586-020-2007-4. Acesso em: 7 ago. 2024.

WOMMACK, K. E.; COLWELL, R. R. **Virioplankton: viruses in aquatic ecosystems. Microbiology and Molecular Biology Reviews**, v. 64, n. 1, p. 69-114, mar. 2000. DOI: 10.1128/MMBR.64.1.69-114.2000. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC98987/. Acesso em: 7 ago. 2024.

WU, S. et al. **DeePhage: distinguishing virulent and temperate phage-derived sequences in metavirome data with a deep learning approach**. Gigascience, 2021. DOI: 10.1093/gigascience/giab056. Disponível em: https://doi.org/10.1093/gigascience/giab056. Acesso em: 7 ago. 2024.

PEREIRA, R.; OLIVEIRA, J.; SOUSA, M. **Bioinformatics and computational tools for next-generation sequencing analysis in clinical genetics**. J Clin Med, v. 9, n. 1, p. 132, 3 jan. 2020. DOI: 10.3390/jcm9010132. Disponível em: https://www.mdpi.com/2077-0383/9/1/132. Acesso em: 7 ago. 2024.

GOODWIN, Sara; McPHERSON, John D.; McCOMBIE, W. Richard. **Coming of age: ten years of next-generation sequencing technologies**. Nat Rev Genet, v. 17, n. 6, p. 333-351, 2016. DOI: 10.1038/nrg.2016.49. Disponível em: https://doi.org/10.1038/nrg.2016.49. Acesso em: 7 ago. 2024.

JEON, Sang Ah et al. **Comparison of the MGISEQ-2000 and Illumina HiSeq 4000 sequencing platforms for RNA sequencing**. Genomics Informatics, v. 17, n. 3, p. e32, 2019.

AMARASINGHE, Sadeep L. et al. **Opportunities and challenges in long-read sequencing data analysis.** Genome Biol, 2020. DOI: 10.1186/s13059-020-1935-5. Disponível em: https://doi.org/10.1186/s13059-020-1935-5. Acesso em: 7 ago. 2024.

TOMMASO, Paolo Di; FLODEN, Evan W.; MAGIS, Cedrik; PALUMBO, Emilio; NOTREDAME, Cedric. **Nextflow: un outil efficace pour l'amélioration de la stabilité numérique des calculs en analyse génomique. Biologie Aujourd'hui**, v. 211, n. 3, p. 233-237, 2017. DOI: 10.1051/jbio/2017029. PMID: 29412134.

CHEN, Shifu; ZHOU, Yonggang; CHEN, Yaru; GU, James. **fastp: an ultra-fast all-in-one FASTQ preprocessor**. Bioinformatics, v. 34, n. 17, p. i884-i890, 2018. DOI: 10.1093/BIOINFORMATICS/BTY560. Disponível em: https://doi.org/10.1093/BIOINFORMATICS/BTY560. Acesso em: 7 ago. 2024.

LI, Dinghua; LIU, Chi-Man; LUO, Ruibang; SADAKANE, Kunihiko; LAM, Tak-Wah. **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, maio 2015. DOI: 10.1093/bioinformatics/btv033. Disponível em: https://doi.org/10.1093/bioinformatics/btv033. Acesso em: 7 ago. 2024.

SHANG, Jiayu; TANG, Xubo; GUO, Ruocheng; SUN, Yanni. **Accurate identification of bacteriophages from metagenomic data using Transformer**. Briefings in Bioinformatics, v. 23, n. 4, p. bbac258, jul. 2022. DOI: 10.1093/bib/bbac258. Disponível em: https://doi.org/10.1093/bib/bbac258. Acesso em: 7 ago. 2024.

NAYFACH, Stephen; CAMARGO, Andre P.; et al. **CheckV assesses the quality and completeness of metagenome-assembled viral genomes.** Nat Biotechnol, v. 39, p. 578-585, 2021. DOI: 10.1038/s41587-020-00774-7. Disponível em: https://doi.org/10.1038/s41587-020-00774-7. Acesso em: 7 ago. 2024.

OLM, Matt; BROWN, Christopher; BROOKS, Brandon; BANFIELD, Jillian. **DRep: A tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication**. The ISME Journal, v. 11, 2017. DOI: 10.1038/ismej.2017.126.

BIN JANG, H.; et al. **Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks.** Nat Biotechnol, v. 37, p. 632-639, 2019. DOI: 10.1038/s41587-019-0100-8. Disponível em: https://doi-org.ez106.periodicos.capes.gov.br/10.1038/s41587-019-0100-8. Acesso em: 7 ago. 2024.

BOURAS, George et al. **Pharokka: a fast scalable bacteriophage annotation tool.** Bioinformatics, v. 39, n. 1, p. btac776, jan. 2023. DOI: 10.1093/bioinformatics/btac776. Disponível em: https://doi.org/10.1093/bioinformatics/btac776. Acesso em: 7 ago. 2024.