UMC UNIVERSIDADE

Revista Científica UMC

Edição Especial PIBIC, outubro 2019 • ISSN 2525-5250

ANÁLISE DE REPLICABILIDADE E REPRODUTIBILIDADE DE UM FLUXO DE TRABALHO PARA ANÁLISES DE MICROBIOMA UTILIZANDO DADOS DE VIDEIRAS APRESENTANDO SINTOMAS DA DOENÇA GALHA-DE-COROA

Marcos Vinicius Yano[1]; Regina Costa de Oliveira[2], Fabiano B. Menegidio[3] Luiz R. Nunes[4]

1. Estudante do curso de Sistemas de Informação; e-mail: rafaelga7@gmail.com
2. Coordenadora Programa do em Biotecnologia; e-mail: reginaco@umc.br
3. Professor Universidade de Mogi das Cruzes; fabiano.menegidio@biology.com.br
4. Professor Universidade de Mogi das Cruzes; e-mail: nunes1212@gmail.com

Área de Conhecimento: **Genética Molecular; Microbiologia.**

Palavras-Chaves: Reprodutibilidade; replicabilidade; microbioma; bioinformática

INTRODUÇÃO

Uma das pedras angulares da metodologia científica é a capacidade de se avaliar criticamente a exatidão das afirmações científicas e as conclusões tiradas por outros pesquisadores em um determinado estudo (CROCKER e COOPER, 2011; SANDVE *et al.*, 2013; PLESSER, 2018). Na última década, eventos classificados como Crises de Reprodutibilidade ganharam importância no meio científico (por exemplo, PASHLER e WAGENMAKERS 2012), por demonstrarem inúmeras falhas na busca de reproduzir e replicar diferentes pesquisas científicas nas mais diversas áreas, incluindo as ciências médicas, comportamentais e da vida (por exemplo, Open Science Collaboration, OSC 2015). Em 2016, Baker relatou em uma pesquisa conduzida pela revista Nature que 90% dos cientistas entrevistados acreditavam que a ciência estava enfrentando tanto uma Crise de Reprodutibilidade quanto de Replicabilidade. Dentre as causas apontadas, mais de 50% dos entrevistados apontaram um baixo poder estatístico e baixa replicabilidade e reprodutibilidade no laboratório como fatores primários. Além disso, fatores como falhas na descrição da metodologia empregada e ausência dos códigos fonte de aplicações e scripts utilizados durante as análises também se mostraram causas para a falha de reprodutibilidade e replicabilidade dos resultados. Mesmo esse sendo um fator extremamente importante para o fazer científico, a noção de disponibilizar informações sobre códigos fonte de aplicações e scripts utilizados durante as análises em conjunto com uma publicação tem sido frequentemente recebida com grande surpresa por diferentes pesquisadores da microbiologia (RAVEL & WOMMACK, 2014). Um dos motivos associados é justamente uma mudança de paradigma, onde os estudos da microbiologia têm abandonado sua característica meramente descritiva e associativa para se apresentar com uma ciência translacional e que manipula um grande conjunto de dados, entrando na era do Big Data. Entre as áreas da microbiologia em que essa mudança se mostra mais presente, destacam-se as Ciências Ômicas, principalmente a Metagenômica e as análises de microbioma.

OBJETIVOS

Pensando nisso, o presente projeto teve como principal objetivo uma análise de replicabilidade de um fluxo de trabalho desenvolvido no laboratório e desenhado para a análise do microbioma central de camundongos caquéticos e saudáveis, testando se o fluxo pode ser generalizado para dados diferentes dos originais. Para a análise de generalização e de replicabilidade foram utilizadas bibliotecas disponíveis no banco de dados European Nucleotide Archive (número de projeto PRJEB12040), provenientes de estudo já publicado por Faist e colaboradores (2016) analisando o microbioma de videiras saudáveis e acometidas pela doença galha-de-coroa. Além disso, realizamos uma análise da reprodutibilidade da

metodologia descrita por Faist e colaboradores (2016), verificando principalmente os detalhes relacionados as análises de bioinformática e se esses poderiam ser reproduzidos conforme descrito no artigo.

METODOLOGIA

A análise conduzida neste projeto utilizou apenas as bibliotecas classificadas como *vineyard* e provenientes de coletas nos troncos de videiros saudáveis e doentes, disponíveis no banco de dados European Nucleotide Archive (número de projeto PRJEB12040) e seus respectivos metadados (MGYS00001337). Estes dados foram submetidos a um fluxo de trabalho desenvolvido no próprio laboratório, sendo ele dividido nas seguintes etapas: análise de qualidade, pré-processamento, processamento e pós-processamento. Durante a etapa de pré-processamento, o *script multiple_join_paired_ends.py* foi utilizado com a finalidade de realizar a junção entre todas as bibliotecas *forward* e *reverse*. Após, o processo de *Splitting Libraries* foi utilizado para renomear o cabeçalho das sequências, realizar o corte de qualidade, concatenar todas as sequências e converter o arquivo para a extensão FASTA. Os *scripts identify_chimeric_seqs.py* e *filter_fasta.py* também foram utilizados para a remoção de todas as sequências identificadas como quimeras e sua posterior filtragem. A etapa de processamento foi responsável pela criação da *OTU table* (unidades taxonômicas operacionais) e a criação do microbioma central. O *script pick_open_reference_otus.py* foi utilizado para a criação da *OTU table,* utilizando o banco de dados do *Greengenes* (DESANTIS, 2006) referência. Após sua criação, a *OTU table* foi submetida a *scripts* de filtragem do pacote QIIME para: (1) remoção de mitocôndrias e cloroplastos contaminantes, (2) descarte de *singletons* e (3) filtragem de OTUs por número de observações. O último *script compute_core_microbiome.py* foi utilizado para criar a *OTU table* final, contendo apenas as *OTUs* presentes em 80% de todas as amostras.

RESULTADOS E DISCUSSÃO

A análise da metodologia de Faist e colaboradores (2016) demonstrou que seus resultados não são reprodutíveis, já que detalhes sobre a metodologia não são disponibilizados no artigo. Detalhes como a versão do banco de dados de referência utilizado, desenho experimental, parâmetros de cada ferramenta que compõe seu fluxo de trabalho e metodologia de clusterização não são fornecidos, impactando que outros pesquisadores realizem novamente as análises descritas. As bibliotecas obtidas no banco de dados European Nucleotide Archive (número de projeto PRJEB12040) renderam um valor absoluto de 4.356.558 leituras. Após toda a etapa de Pré-Processamento, observamos a permanência de 2.035.218 sequências (46,7%) dos dados provenientes das bibliotecas de videira para a realização da etapa posterior de Processamento. Em contraste, notamos a permanência de 10.855.262 sequências (21,6%) dos dados provenientes de camundongos para as etapas posteriores. Essa diferença se mostrou pela qualidade das amostras e o número de quimeras identificadas em ambas as bibliotecas. Graças ao processo de joined pair-end, que naturalmente reduz a quantidade de pares de base quase pela metade, e pelos filtros de qualidade agregados baseado no valor de Q20, torna-se esperado uma alta redução de sequências nessa etapa da análise (CAPORASO et al., 2010;). Essas mudanças na proporção de sequências fornece uma melhor qualidade nos dados gerados e reduz a clusterização equivocada de sequências provenientes de quimeras geradas no processo de sequenciamento. Posteriormente, essas sequencias foram submetidas ao processo de clusterização e obtivemos os seguintes valores respectivamente nos dados de videira e camundongos: 2.031.465 sequências (46.6%) e 10.637.921 sequências (21.2%). Após a criação da tabela de OTUs iniciais, a etapa de filtragem gerou uma grande queda no número de sequências para os dados de videira, passando de 2.031.465 sequências (46.6%) para apenas 207.519 sequências (4.7%). Uma análise mais detalhada dos resultados nos mostrou que essa redução acentuada de sequências não gerou um impacto no número de

observações de OTU no conjunto de dados. Enquanto visualizávamos 2.463 OTUs com 2.031.465 sequências, permanecemos com 2.130 observações com 207.519 sequências. Ao avaliar as sequências e OTUs perdidas durante o processo de filtragem, notamos que sua grande maioria foi identificada como cloroplastos putativos, 279 OTUs com 1.246.743 sequências. Além disso, 54 OTUS com 577.203 sequências foram identificadas como mitocôndrias putativas. A existência de sequências de cloroplastos e mitocôndrias são esperadas em dados de microbioma, mas a filtragem (ou não filtragem) de 1.823.946 sequências, correspondendo 41.8% das 2.031.465 sequências utilizadas durante o processo de clusterização pode gerar erros ou falsas conclusões quando as análises estatísticas forem realizadas posteriormente. Para os dados de camundongos não observamos esse alto valor de mitocôndrias e cloroplastos, sendo que esses valores correspondiam a 21.1% dos dados obtidos na tabela de OTUs originais.

CONCLUSÕES

Com base nos resultados obtidos, podemos verificar a importância dos conceitos de repetibilidade, reprodutibilidade e replicabilidade no desenvolvimento científico, incluindo as áreas de Microbioma e Bioinformática. Nossos resultados demonstram que o processo de repetibilidade e reprodutibilidade não são suficientes para o fazer científico, sendo que promover a replicabilidade dos resultados se torna de extrema importância para a validação dos resultados obtidos anteriormente. Demonstramos também que a disponibilidade de todos os metadados relacionados as bibliotecas, versões e parâmetros dos softwares utilizados é essencial para que outros grupos de pesquisas possam replicar os resultados apresentados. Também demonstramos a importância de testes internos para garantir que fluxos de trabalho possam ser reproduzidos com os dados brutos disponibilizados, além de testes para validar se um fluxo de trabalho pode ser replicado e generalizado para diferentes bibliotecas. Acreditamos que o presente trabalho colaborou na promoção dos conceitos relatados e auxiliou nas análises relacionadas a Crise de Reprodutibilidade e Replicabilidade na Ciência.

REFERÊNCIAS

BAKER, M. 1,500 Scientists Lift the Lid on Reproducibility, **Nature**, 533(7604): 452–454. 2016.

CAPORASO, J. G.; KUCZYNSKI, J.; STOMBAUGH, J.; BITTINGER, K.; BUSHMAN, F. D.; COSTELLO, E. K.; FIERER, N.; PENA, A. G.; GOODRICH, J. K.; GORDON, J. I.; HUTTLEY, G. A. QIIME allows analysis of high-throughput community sequencing data. Nature methods. 7(5):335, 2010.

CROCKER J, COOPER ML Addressing scientific fraud. **Science** 334: 1182, 2011.

DESANTIS, TZ et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. **Appl Environ Microb.** 72(7): 5069-5072, 2006.

DHARIWAL, A. et al. MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. **Nucleic acids research** 45.W1 W180-W188 2017.

FAIST, H. et al. Grapevine (Vitis vinifera) crown galls host distinct microbiota. **Appl. Environ. Microbiol.** 82.18 5542-5552, 2016.

LANGILLE, MG. et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. **Nat Biotechnol.** v.31, n.9, p.814-21, 2013.

PASHLER, H. & WAGENMAKERS E. Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence, **Perspectives on Psychological Science**, 7(6): 528–530. 2012

PLESSER HE. Reproducibility vs. Replicability: A Brief History of a Confused Terminology. **Front Neuroinform.** 2018;11:76. Published 2018.

RAVEL J, WOMMACK KE. All hail reproducibility in microbiome research. **Microbiome** 2:8. 2014.

SANDVE GK et al. Ten Simple Rules for Reproducible Computational Research. **PLOS Computational Biology** 9(10): e1003285. 2013