APLICAÇÃO DE PRINCÍPIOS FAIR E ORQUESTRADORES DE FLUXOS DE TRABALHO NO CONTEXTO DE ANÁLISES DE DADOS

Maria do Ceu Pontes Vieira[1]; Marianna Pereira Silva Ramalho[1]; Yasmin Yngrid Mendes de Brito[1]; Gabriela Pessoa Lima de Souza Medeiros[1]; Marcel da Câmara Ribeiro-Dantas[2,3] (Dr.)
[1]Curso de Medicina, Universidade Potiguar.
[2]Programa de Pós-Graduação em Biotecnologia, Universidade Potiguar.
[3]Programa de Pós-Graduação em Administração, Universidade Potiguar.
e-mail: marcel.camara@ulife.com.br

RESUMO

Softwares de orquestração de fluxos de trabalho, como o Nextflow, têm revolucionado a análise de dados, particularmente na saúde, ao simplificar a integração de ferramentas bioinformáticas em infraestruturas variadas. Em paralelo, os princípios FAIR têm adquirido relevância acentuada nesse mesmo contexto. Este estudo aborda a relevância desses princípios e tecnologias no gerenciamento de dados complexos e na reprodutibilidade de pesquisas, destacando sua aplicação na análise de genomas clínicos, através de uma revisão narrativa da literatura. Os achados demonstram redução significativa no tempo de execução e aumento na eficiência dos fluxos de trabalho, com impacto positivo na medicina personalizada. Concluímos que essas ferramentas e princípios são indispensáveis para a ciência moderna e apresentam potencial para transformar práticas de análise de dados em diversas áreas.

PALAVRAS-CHAVE: Orquestradores de fluxo de trabalho, Bioinformática, FAIR.

INTRODUÇÃO

A análise de dados na saúde enfrenta desafios crescentes devido ao volume e complexidade dos dados gerados, como genomas, exames clínicos e imagens médicas. Orquestradores de fluxo de trabalho, como Nextflow e Snakemake permitem o gerenciamento eficiente de pipelines complexos, facilitando a integração de ferramentas heterogêneas e promovendo a reprodutibilidade. Este estudo analisa a importância desses sistemas, com ênfase no impacto na medicina personalizada.

MÉTODO

Este trabalho trata-se de uma revisão narrativa da literatura. Foi construído através da leitura e interpretação dos artigos selecionados e realizado utilizando a base de dados PubMed, com os descritores: “*workflow management system”* ou “*workflow management systems”* e *scalability* ou *reproducibility* ou *portability* ou *FAIR*, sendo

“e” e “ou” operadores lógicos. Foram incluídos artigos publicados de 2014 a 2024, em inglês, e com o artigo completo disponível gratuitamente, totalizando 53 publicações. Foram excluídos trabalhos que não utilizaram nenhuma tecnologia para orquestração de fluxo de trabalho, que não tratavam de análise de dados, cujos sistemas foram criados em torno de um único fluxo de trabalho ou tipo de fluxo de trabalho, ou que as pesquisas abordavam os temas indiretamente ou superficialmente. Após a aplicação destes critérios de exclusão, o número final foi de 44 trabalhos.

Além dos trabalhos encontrados na revisão, outras publicações foram identificadas e analisadas durante a leitura dos trabalhos encontrados.

RESULTADOS E DISCUSSÕES

Frente ao crescente número de *workflow management systems*, Jackson et al (2021) levantaram um conjunto de critérios relevantes para investigar qual a melhor ferramenta para orquestrar o RoboViz, um pacote para extrair informações biológicas de dados de perfis de ribossomos para ajudar no avanço da compreensão da síntese de proteínas. O estudo concluiu que o Nextflow seria a melhor escolha, o que parece estar alinhado com a decisão de outros pesquisadores que, reconhecendo as barreiras para se desenvolver análises reprodutíveis, optaram por utilizar esses sistemas. Isso foi observado na análise bibliográfica de Langer et al (2024), apontando o Nextflow como o sistema que mais cresceu em número de citações em 2023 comparado com anos anteriores. Spišaková et al investigou a execução do nf-core/sarek, um pipeline orquestrado com Nextflow, em diversos tipos de infraestruturas diferentes, comparando os recursos utilizados, dentre outras métricas.

Em um estudo por Grayson et al (2023), os autores avaliaram a replicabilidade e reprodutibilidade de pipelines publicamente disponíveis e orquestrados pelo Snakemake e pelo Nextflow. No subconjunto de fluxos de trabalho selecionados pelos autores, os do Nextflow tiveram uma taxa de reprodutibilidade sem falhas muito maior, 51%, em comparação aos fluxos de trabalho do Snakemake, 11%. Os autores também mencionaram uma melhor curadoria no repositório de pipelines do

Nextflow, o nf-core, embora ambos os repositórios evidenciem um esforço de oferecer análises curadas e reprodutíveis frente a essa demanda.

Comparando-se aos métodos tradicionais, observou-se maior reprodutibilidade, essencial para a validação de descobertas clínicas. Esses resultados corroboram estudos prévios que apontam a relevância dos orquestradores no avanço da bioinformática e da saúde digital.

A pandemia ressaltou o papel da bioinformática no combate a surtos virais. Pipelines habilitados pelo Nextflow, como nf-core/viralrecon, ncov2019-artic-nf e Elan, impulsionaram o sequenciamento e análise do SARS-CoV-2, junto ao controle de qualidade e a agregação de dados em laboratórios globais. O ASPIcov é um outro exemplo de pipeline desenvolvido visando fornecer uma análise rápida, confiável e completa das amostras de SARS-CoV-2. Afiahayati et al realizaram um estudo com base em um método de captura de hibridação capaz de capturar vírus respiratórios direcionados, incluindo o SARS-CoV-2. A partir disso surgiram dois pipelines para análise desses dados, o ‘Fast Pipeline' e o ‘Normal Pipeline'. Além disso, fluxos de trabalho colaborativos de código aberto, ferramentas em contêineres e plataformas de orquestração escaláveis garantiram análises eficientes e padronizadas. Pipelines centrais processaram milhões de genomas, alcançando anos de trabalho em dias, cruciais para vigilância em tempo real e rastreamento de variantes. Essas inovações facilitaram insights genômicos rápidos, capacitando respostas de saúde pública em meio a desafios globais sem precedentes. Esses avanços, potencializados pelo uso do Nextflow, permitiram a identificação de diversas variantes do SARS-CoV-2, como a Alpha e Delta.

As abordagens adotadas incluem o uso de pipelines de análise de dados reprodutíveis e fluxos de trabalho de análise de sequência de genoma, usando tecnologias como o gerenciador de fluxo de trabalho Nextflow. O bactopia, por exemplo, é um pipeline que utiliza o Nextflow com o objetivo de fornecer análises genômicas comparativas eficientes para espécies ou gêneros bacterianos. Este pipeline é particularmente valioso para vigilância genômica de saúde pública. Ele permite a análise rápida e abrangente de genomas bacterianos, o que é crucial para monitorar e responder a surtos, rastrear a disseminação de patógenos e detectar

ameaças emergentes, indo além do exemplo recente mais impactante que era a vigilância genômica de um vírus, o SARS-CoV-2.

CONCLUSÕES

Os orquestradores de fluxo de trabalho, como o Nextflow, são ferramentas indispensáveis na análise de dados de saúde, promovendo eficiência, reprodutibilidade e escalabilidade. Esta revisão reforça seu papel transformador na medicina personalizada, contribuindo para a análise de grandes volumes de dados clínicos de forma robusta, eficiente e reprodutível.

REFERÊNCIAS

Di Tommaso, Paolo, et al. "**Nextflow enables reproducible computational workflows**". Nature Biotechnology 35.4 (2017): 316-319.

Langer, Bjorn E., et al. "**Empowering bioinformatics communities with Nextflow and nf-core**". bioRxiv (2024): 2024-05.

Floden, Evan. "**Genetic Sequencing Will Enable Us To Win The Global Battle Against COVID-19**". https://www.bio-itworld.com/news/2021/11/05/genetic-sequencing-will-enable-us-to-win-the-global-battle-against-covid-19. Acessado em: 19 de Novembro de 2024.

Tilloy, Valentin; Cuzin, Pierre; Leroi, Laura; Guérin, Emilie; Durand, Patrick; Alain, Sophie (2022). "**ASPICov: An automated pipeline for identification of SARS-Cov2 nucleotidic variants**". PLOS ONE. 17 (1): e0262953. Bibcode:2022PLoSO..1762953T.

Petit, Robert A.; Read, Timothy D. (2020). "**Bactopia: A Flexible Pipeline for Complete Analysis of Bacterial Genomes**". mSystems. 5 (4). doi:10.1128/mSystems.00190-20.

Brandt, Christian; Krautwurst, Sebastian; Spott, Riccardo; Lohde, Mara; Jundzill, Mateusz; Marquet, Mike; Hölzer, Martin (2021). "**Pore Cov-An Easy to Use, Fast,**

and Robust Workflow for SARS-CoV-2 Genome Reconstruction via Nanopore Sequencing". Frontiers in Genetics. 12: 711437. doi:10.3389/fgene.2021.711437.

Afiahayati; Bernard, Stefanus; Gunadi; Wibawa, Hendra; Hakim, Mohamad Saifudin; Marcellus; Parikesit, Arli Aditya; Dewa, Chandra Kusuma; Sakakibara, Yasubumi (2022). "**A Comparison of Bioinformatics Pipelines for Enrichment Illumina Next Generation Sequencing Systems in Detecting SARS-CoV-2 Virus Strains**". Genes. 13 (8): 1330. doi:10.3390/genes13081330.

Ahmed, A.E., Allen, J.M., Bhat, T. et al. "**Design considerations for workflow management systems use in production genomics research and the clinic**". Sci Rep 11, 21680 (2021). https://doi.org/10.1038/s41598-021-99288-8

Vieira, Maria do Céu Pontes; Medeiros, Gabriela Pessoa Lima de Souza; Ramalho, Marianna Pereira Silva; Brito, Yasmin Yngrid Mendes de; Ribeiro-Dantas, Marcel da Câmara (2024). "**Estudo Genômico de Neoplasia Mamária com Base na Utilização de Pipeline**". III Congresso Nacional Multidisciplinar de Oncologia.

Ramalho, Marianna Pereira Silva; Vieira, Maria do Céu Pontes; Brito, Yasmin Yngrid Mendes de; Medeiros, Gabriela Pessoa Lima de Souza; Ribeir-Dantas, Marcel da Câmara (2024). "**Retinoblastoma Infantil: Diagnóstico Precoce e Análise dos Benefícios do uso de Nanopartículas no Tratamento**". III Congresso Nacional Multidisciplinar de Oncologia.

Medeiros, Gabriela Pessoa Lima de Souza; Brito, Yasmin Yngrid Mendes de; Vieira, Maria do Céu Pontes; Ramalho, Marianna Pereira Silva; Ribeiro-Dantas, Marcel da Câmara (2024). "**Orquestradores de Luxo de Trabalho na Análise Genômica do HPV: Benefícios e Desafios**". III Congresso Nacional Multidisciplinar de Oncologia.

Jackson M, Kavoussanakis K, Wallace EWJ (2021) "**Using prototyping to choose a bioinformatics workflow management system**". PLOS Computational Biology 17(2): e1008622. https://doi.org/10.1371/journal.pcbi.1008622

Spišaková, Viktória; Hejtmánek, Lukáš; Hynšt, Jakub (2023) “**Nextflow in Bioinformatics: Executors Performance Comparison Using Genomics Data**”. Future Generation Computer Systems, Volume 142, 2023, Pages 328-339, ISSN 0167-739X, https://doi.org/10.1016/j.future.2023.01.009.

Grayson, Samuel; Marinov, Darko; Katz, Daniel S. and Milewicz, Reed. 2023. “**Automatic Reproduction of Workflows in the Snakemake Workflow Catalog and nf-core Registries**”. In Proceedings of the 2023 ACM Conference on Reproducibility and Replicability (ACM REP '23). Association for Computing Machinery, New York, NY, USA, 74–84. https://doi.org/10.1145/3589806.3600037.

FOMENTO

M.C.R.D. recebeu financiamento através dos editais 35/2023 e 65/2024 do Instituto Ânima.