25º enancib
Encontro Nacional de Pesquisa e Pós-graduação em Ciência da Informação
Informação, Decolonialidade e Direitos Difusos para o Desenvolvimento Sustentável
3 a 7 de novembro de 2025 - Rio de Janeiro

XXV ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – XXV ENANCIB

GT 08 – Dados, Informação e Tecnologia

GESTÃO FAIR (*Findable, Accessible, Interoperable, Reusable*) DE DADOS DE PESQUISA MULTIÔMICOS: Análise comparativa de ferramentas e infraestrutura

FAIR MANAGEMENT OF MULTI-AMERICAN RESEARCH DATA: Comparative analysis of tools and infrastructure

Fábio Bernardo da Silva – Fundação Osvaldo Cruz (FIOCRUZ)
Anderson Silva de Araujo – Fundação Oswaldo Cruz (FIOCRUZ)
Viviane Santos de Oliveira Veiga - Fundação Osvaldo Cruz (FIOCRUZ)

Modalidade: Trabalho Completo

Resumo: a crescente geração de dados multiômicos, especialmente genômicos, desafia a gestão, análise e reúso eficiente dessas informações na pesquisa biomédica. Nesse cenário, os princípios FAIR (*Findable, Accessible, Interoperable* e *Reusable*) destacam-se como diretrizes essenciais para garantir reprodutibilidade e transparência científica. Este artigo analisa comparativamente ferramentas e infraestruturas de código aberto voltadas à gestão FAIR de metadados e à análise de dados multiômicos. A metodologia incluiu uma revisão descritiva e comparativa de plataformas, baseada em documentação técnica, artigos e diretrizes internacionais. Foram avaliadas: *Galaxy, Nextflow* (com nf-core), *FAIRDOMHub*, *FAIR Data Cube*, *ISA-tools* e *Vantage6*. A análise mostrou que todas as plataformas oferecem contribuições relevantes, embora com escopos e níveis de aderência FAIR distintos. *Galaxy* e *Nextflow* destacam-se na bioinformática e reprodutibilidade, com pipelines robustos. FAIR Data Cube e Vantage6 avançam no tratamento ético de dados sensíveis, permitindo análise federada sem compartilhamento entre instituições. A adoção de padrões como ISA-Tab, RDF (*Resource Description Framework*), ontologias OBO (*Open Biomedical Ontologies*) e identificadores persistentes aumenta a interoperabilidade e facilita a reutilização eficaz. As Infraestruturas abertas e modulares, alinhadas a arquiteturas *FAIR-at-source*, não só cumprem exigências regulatórias, mas aceleram a ciência translacional em contextos interdisciplinares e multicêntricos. Conclui-se que ecossistemas computacionais abertos e *FAIR* são essenciais para o avanço da pesquisa multiômica em saúde, demandando investimentos contínuos em infraestrutura, padronização semântica e capacitação técnica. A integração dessas soluções promove não apenas a gestão eficiente de dados complexos, mas também a colaboração global, reforçando a confiabilidade e o impacto das descobertas científicas.

Palavras-chave: princípios *FAIR*; dados multiômicos; metadados; interoperabilidade; bioinformática.

Abstract: the increasing generation of multiomics data, especially genomics, challenges the efficient management, analysis and reuse of this information in biomedical research. In this scenario, the FAIR (Findable, Accessible, Interoperable and Reusable) principles stand out as

essential guidelines to ensure reproducibility and scientific transparency. This article comparatively analyzes open source tools and infrastructures aimed at FAIR metadata management and multiomics data analysis. The methodology included a descriptive and comparative review of platforms, based on technical documentation, articles and international guidelines. The following platforms were evaluated: Galaxy, Nextflow (with nf-core), FAIRDOMHub, FAIR Data Cube, ISA-tools, and Vantage6. The analysis showed that all platforms offer relevant contributions, although with different scopes and levels of FAIR adherence. Galaxy and Nextflow stand out in bioinformatics and reproducibility, with robust pipelines. FAIR Data Cube and Vantage6 advance the ethical treatment of sensitive data, allowing federated analysis without sharing between institutions. The adoption of standards such as ISA-Tab, RDF, OBO ontologies and persistent identifiers increases interoperability and facilitates effective reuse. Open and modular infrastructures, aligned with FAIR-at-source architectures, not only meet regulatory requirements but also accelerate translational science in interdisciplinary and multicenter contexts. It is concluded that open and FAIR computational ecosystems are essential for the advancement of multiomics research in health, demanding continuous investments in infrastructure, semantic standardization and technical capacity. The integration of these solutions promotes not only the efficient management of complex data, but also global collaboration, reinforcing the reliability and impact of scientific discoveries.

Keywords: FAIR principles; multi-omic data; metadata; interoperability; bioinformatics**.**

1 INTRODUÇÃO

Com o avanço das tecnologias de sequenciamento de nova geração NGS (*Next-Generation Sequencing*) ou Sequenciamento de Nova Geração), proteômica e metabolômica, os dados multiômicos tornaram-se centrais na pesquisa biomédica (Hasin; Sagi; Marian, 2017). No entanto, a complexidade e o volume desses dados impõem desafios relacionados à sua gestão, compartilhamento e análise. Os princípios FAIR, propostos em 2016, surgem como resposta a essa demanda, orientando a ciência para maior transparência, acessibilidade, interoperabilidade e reprodutibilidade (Wilkinson *et al*., 2016).

Este artigo tem como objetivo analisar comparativamente ferramentas e infraestruturas de código aberto voltadas à gestão FAIR de metadados e à análise de dados multiômicos. A pesquisa busca responder à seguinte questão: quais são as funcionalidades, alinhamento com os princípios FAIR e potencial de integração em ecossistemas reprodutíveis, escaláveis e seguros das ferramentas e infraestruturas digitais de código aberto para gestão FAIR de metadados e análise de dados multiômicos na pesquisa científica?

A justificativa para este estudo reside na crescente necessidade de otimizar a gestão e o reúso de dados complexos na pesquisa biomédica. A adoção de ecossistemas

computacionais abertos e FAIR é crucial para o avanço da pesquisa multiômica em saúde, demandando investimentos contínuos em infraestrutura, padronização semântica e capacitação técnica. A integração dessas soluções promove não apenas a gestão eficiente de dados complexos, mas também a colaboração global, reforçando a confiabilidade e o impacto das descobertas científicas.

As Ferramentas digitais compatíveis com FAIR são essenciais para estruturar *pipelines* analíticos reprodutíveis e garantir que dados possam ser reutilizados por diferentes grupos e em diferentes contextos (Sansone *et al*., 2019). O presente estudo realiza uma análise comparativa de ferramentas e infraestruturas que atendem a esses critérios no contexto da ciência multiômica.

2 REVISÃO DA LITERATURA

A Ciência Aberta (*Open Science*), entendida como movimento transformador que defende o conhecimento científico como bem público acessível, transparente e reutilizável (UNESCO, 2022), fundamenta-se em pilares como reprodutibilidade, reuso de dados e princípios FAIR. Na área de genômica, esses elementos são particularmente críticos: frameworks que adotam identificadores persistentes, modelos de metadados padronizados e catálogos semânticos asseguram a localização, acesso e rastreabilidade de dados, conforme evidenciado na análise comparativa. Tal infraestrutura não apenas viabiliza a citação adequada, atribuindo créditos aos pesquisadores, como também garante a reprodutibilidade analítica, condição essencial para o avanço científico.

Ao implementar mecanismos rigorosos de controle de acesso e proteção, tais frameworks promovem uma Ciência Aberta responsável (Van Vliet; Moore, 2016), alinhando-se aos princípios de ética, transparência e equidade. Esta sinergia entre abertura e segurança, quando integrada à etapa final da análise, demonstra como a operacionalização dos pilares FAIR não apenas dialoga com os fundamentos da Ciência Aberta, mas também otimiza a governança de dados genômicos, potencializando colaborações e assegurando rigor metodológico (Azevedo; Mendonça, 2024; Haddaway, 2018).

2.1 Princípios FAIR: Contextos e Fundamentos

A formulação dos princípios FAIR ocorreu em 2016 por meio de articulação entre pesquisadores e iniciativas internacionais como *GO FAIR, FORCE11 e EOSC (European Open*

Science Cloud) (Wilkinson et al., 2016). Ao contrário de abordagens voltadas apenas à abertura dos dados, os princípios FAIR enfatizam a estruturação semântica e a reutilização efetiva por máquinas e humanos (Mons *et al*., 2020). Sua implementação foi incentivada por agências como *ELIXIR, NIH* e Comissão Europeia (Jacobsen *et al*., 2020).

Diferentemente de iniciativas anteriores focadas apenas na abertura (open data), os princípios FAIR enfatizam a necessidade de qualidade semântica, estrutura padronizada e metadados ricos que possibilitem o reencontro, a integração e o reúso de dados científicos de forma automatizada. Cada princípio é desdobrado em critérios técnicos específicos: por exemplo, "Findable" exige uso de identificadores persistentes como DOI (*Digital Object Identifier*) e ORCID (*Open Researcher and Contributor ID*) e catálogos indexáveis; "Accessible" implica que os dados estejam disponíveis via protocolos abertos, mesmo quando com restrições de acesso; "Interoperable" pressupõe uso de ontologias formalizadas e vocabulários compartilhados; e "Reusable" requer licenciamento claro e proveniência detalhada.

A adoção dos princípios FAIR ganhou força especialmente em áreas com produção massiva de dados, como as ciências ômicas, a saúde digital e a astronomia. Organizações como a *ELIXIR* (para ciências da vida), o *NIH* (nos EUA) e a Comissão Europeia passaram a exigir que projetos financiados implementassem estratégias FAIR desde o desenho experimental até a publicação dos dados. Nesse cenário, surgiram diversas ferramentas e infraestruturas de código aberto como *FAIRDOMHub, FAIR Data Point e FAIR Data Cube*, para operacionalizar esses princípios, tornando-os aplicáveis na prática científica cotidiana.

A criação dos princípios FAIR representa, portanto, um marco conceitual e técnico na governança de dados científicos, promovendo uma cultura orientada à transparência, reprodutibilidade e colaboração em larga escala, especialmente em contextos interdisciplinares e multicêntricos.

2.2 *FAIR Data Cube (FDCUBE)*

Desenvolvido pela *X-omics Initiative* (Países Baixos), o *FAIR Data Cube* é um *framework* de código aberto que une *metadata FAIR-at-source*, descoberta semântica e análise federada de dados multiômicos sensíveis (Witjes *et al*., 2022). Sua arquitetura combina a *FAIR Data Station, FAIR Data Point (FDP) e Vantage6*, implementando o conceito de computação federada conforme os princípios do *Personal Health Train* (Deist *et al*., 2020). Além disso, o *FDCube* utiliza *GraphDB* como *triplestore*, adota *Phenopackets* para descrição de fenótipos

humanos e já é distribuído como pacote "*in-a-box*" na *SURF Research Cloud*. No projeto *Trusted World of Corona*, a plataforma permitiu a análise federada sem deslocamento de dados brutos (Witjes et al., 2022).

Privacidade e governança dos dados, interoperabilidade nativa com ontologias padronizadas, escalabilidade via *Kubernetes* e *Nextflow* e transparência com versionamento e preservação de logs tornam o FDCube uma solução robusta (Deist et al., 2020).

A configuração do FDCube ainda demanda conhecimento técnico elevado e sua documentação encontra-se majoritariamente em inglês, limitando a adoção em países lusófonos (Witjes *et al*., 2022).

2.3 Dados Genômicos: Volume, Complexidade e Aplicações na Saúde

Os dados genômicos compõem um dos maiores e mais complexos conjuntos de informações biomédicas da atualidade. A partir do Projeto Genoma Humano (2001), que decodificou cerca de 3 bilhões de pares de bases do DNA humano ao custo de quase 3 bilhões de dólares, o avanço nas tecnologias de sequenciamento de nova geração (NGS) reduziu drasticamente o custo e o tempo necessários para obter genomas completos. Atualmente, é possível sequenciar um genoma humano por menos de 500 dólares em poucas horas, gerando entre 100 e 200 *gigabyte*s de dados brutos por indivíduo (Wilkinson et al, 2016). Estima-se que, até 2030, mais de 100 milhões de genomas humanos terão sido sequenciados apenas no contexto clínico, com produção global de dados genômicos ultrapassando *exabytes* anuais (Sansone *et al.*, 2021).

A riqueza dos dados genômicos permite aplicações estratégicas na medicina personalizada, na diagnose de doenças raras e genéticas, na oncologia de precisão e na farmacogenômica, além de fornecer subsídios para intervenções preditivas e preventivas na saúde pública. Em câncer, por exemplo, a análise de mutações somáticas orienta a escolha de terapias-alvo e imunoterapias. Em doenças raras, o sequenciamento de exoma ou genoma inteiro acelera o diagnóstico diferencial. Além disso, consórcios como o 100,000 *Genomes Project* (Reino Unido) e o *Genomic Data Commons* (NIH/EUA) promovem repositórios abertos para reúso de dados genômicos em larga escala, integrando-os a dados clínicos e ambientais.

Contudo, os desafios são significativos. A governança ética e legal dos dados genômicos, por serem altamente identificáveis e sensíveis exige mecanismos robustos de consentimento, anonimização e controle de acesso. Do ponto de vista computacional, a

análise e armazenamento requerem infraestrutura de alto desempenho e soluções escaláveis que atendam aos princípios FAIR, possibilitando reúso seguro e interoperável. Nesse contexto, ferramentas abertas como *Galaxy, Nextflow*, e plataformas como o *FAIR Data Cube* têm se consolidado como alternativas viáveis para lidar com o volume e a sensibilidade dos dados genômicos na pesquisa translacional e na prática clínica.

2.4 Infraestrutura das análises genômicas: Componentes e funções

A análise genômica requer uma infraestrutura complexa, composta por diferentes camadas de tecnologias que garantem desde a aquisição dos dados até sua interpretação e reúso. Em linhas gerais, essa infraestrutura pode ser dividida em cinco componentes principais: aquisição de dados, armazenamento, processamento, anotação e interpretação, e compartilhamento seguro.

A aquisição de dados envolve plataformas de sequenciamento de nova geração (NGS), como *Illumina, Oxford Nanopore* e PacBio, responsáveis por produzir leituras de DNA ou RNA em alta velocidade. Esses dados brutos gerados em formato FASTQ contêm bilhões de pares de bases e são o ponto de partida para análises posteriores.

Na camada de armazenamento, são utilizados servidores locais de alta capacidade, clusters computacionais ou serviços de nuvem (como *AWS, Google Cloud e EGA*), que precisam ser compatíveis com formatos padronizados (FASTQ, BAM, VCF) e dotados de mecanismos de compressão, criptografia e versionamento.

O processamento é conduzido por pipelines bioinformáticos que alinham leituras (com ferramentas como *BWA* ou *STAR*), realizam chamadas de variantes (GATK, *FreeBayes*) e quantificam expressão gênica (*featureCounts, Salmon*). Aqui, entram também orquestradores de workflows como *Nextflow, Snakemake* e *Galaxy*, que garantem reprodutibilidade, paralelização e integração com ambientes *Docker* ou *Singularity*.

A etapa de anotação e interpretação adiciona contexto funcional às variantes genéticas identificadas, utilizando bancos como *Ensembl, ClinVar* e *dbSNP*. Ferramentas como *ANNOVAR* e *VEP* mapeiam mutações a genes, efeitos biológicos e relevância clínica.

Por fim, o compartilhamento seguro dos dados exige repositórios compatíveis com os princípios FAIR e políticas de acesso controlado, como o *FAIR Data Point*, o *dbGaP* ou o *European Genome-phenome Archive* (EGA). Mecanismos de análise federada, como o

Vantage6, vêm sendo incorporados para permitir a computação sobre dados sensíveis sem expô-los diretamente.

Essa arquitetura modular garante que as análises genômicas sejam tecnicamente viáveis, eticamente seguras e cientificamente reprodutíveis, com crescente adesão às boas práticas de governança de dados em saúde.

3 OBJETIVO

O objetivo deste trabalho é analisar comparativamente as ferramentas e infraestruturas digitais de código aberto voltadas para a gestão FAIR de metadados e para a análise de dados multiômicos na pesquisa científica, com ênfase na identificação de suas funcionalidades, alinhamento com os princípios FAIR e potencial de integração em ecossistemas reprodutíveis, escaláveis e seguros. Alinhados aos pilares da ciência aberta.

4 PROCEDIMENTOS METODOLÓGICOS

Este estudo caracteriza-se como uma pesquisa bibliográfica e documental, realizada na área da Biomedicina. A abordagem adotada foi qualitativa, descritiva e comparativa, focando em ferramentas e infraestruturas digitais empregadas na gestão FAIR de metadados e na análise de dados multiômicos, com especial atenção à genômica. As plataformas selecionadas para análise são reconhecidas na literatura científica e adotadas em consórcios internacionais de pesquisa biomédica, como *Galaxy, Nextflow, FAIRDOMHub, FAIR Data Cube e Vantage* (Di Tommaso, 2017). A seleção baseou-se em três critérios principais: (i) alinhamento explícito com os princípios FAIR (*Findable, Accessible, Interoperable, Reusable*); (ii) aplicação documentada em estudos genômicos ou multiômicos; e (iii) disponibilidade pública com documentação e suporte comunitário.

A coleta de dados foi realizada em março de 2025, por meio de revisão de literatura técnico-científica em bases como *PubMed*, *Scopus* e *Google Scholar*, utilizando os descritores "*FAIR data*", "*genomic analysis infrastructure*", "*open source bioinformatics tools*" e "*multi-omics platforms*". Também foram incluídos documentos técnicos de organizações como *ELIXIR, GO FAIR e NIH*, além da análise direta de manuais de usuário, repositórios *GitHub* e ambientes de demonstração das ferramentas.

A etapa de análise consistiu em uma comparação estruturada das plataformas selecionadas, categorizadas em função de sua posição no fluxo de trabalho multiômico

(captura de metadados, armazenamento, processamento, análise, publicação e reúso). Para cada ferramenta, foram avaliados: (i) componentes tecnológicos; (ii) aderência a padrões internacionais de interoperabilidade (ex: ISA-Tab, RDF, OBO Foundry); (iii) presença de mecanismos de versionamento e identificação persistente; (iv) capacidade de integração com pipelines de análise; e (v) compatibilidade com dados sensíveis via execução local ou federada.

5 RESULTADOS E ANÁLISES

A análise dos dados coletados foi realizada entre março e maio de 2025, com apoio de ambientes computacionais locais e repositórios públicos. Os resultados estão sistematizados em uma tabela comparativa, apresentada abaixo, permitindo destacar os pontos fortes, limitações e complementaridades entre as ferramentas analisadas. A categorização da aderência aos princípios FAIR seguiu os parâmetros definidos por Wilkinson et al. (2016) e adaptados por Mons *et al*. (2020) para aplicações em ciências da vida.

A seguir, apresenta-se um quadro comparativo que sintetiza as características das ferramentas e infraestruturas analisadas, facilitando a visualização de pontos fortes, limitações e exemplos de uso. Os critérios utilizados para a comparação foram desenvolvidos e adaptados a partir das recomendações de Wilkinson *et al*. (2016) e Mons *et al*. (2020), conforme descritos a seguir:

(i) camada funcional no fluxo multiômico - compreende a etapa de interpretação biológica dos dados integrados, buscando compreender o significado funcional das alterações observadas nas diferentes “ômicas” (genômica, proteômica, transcriptômica e metabolômica);

(ii) componentes tecnológicos - refere-se as plataformas e ferramentas de softwares, infraestrutura de hardware que comportam e auxiliam a análise funcional e biológica dos dados, transformando dados brutos ômicos em *insights* biológicos e acionáveis por máquina;

(iii) aplicação típica - Abrange a concepção de experimentos desde pesquisas básicas até soluções clínicas e industriais;

(iv) grau de aderência aos princípios FAIR – compreende a aderência dos dados científicos as diretrizes internacionais para gestão de dados, visando torná-los; *Findable* (localizáveis), *Accessible* (Acessíveis), *Interoperable* (Interoperáveis) e *Reusable* (Reutilizáveis);

(v) pontos fortes – São os elementos que faz essas abordagens serem revolucionárias como: Integração de dados, inovação tecnológica, precisão e personalização e sustentabilidade em saúde;

(vi) limitações - Engloba a complexidade técnica, barreiras éticas e sociais, viés e qualidade de dados, os desafios de implementação teóricos e práticos.

QUADRO 1 – Estruturas de software e suas aplicações na gestão fair de metadados e na análise de dados multiômicos

Ferramenta / Infraestrutura	Camada funcional no fluxo multiômico	Principais componentes / tecnologias	Aplicação típica	Adesão FAIR†	Pontos fortes	Limitações	Exemplo recente
ISA-Tools	Captura e modelagem de metadados	*ISAcreator, ISA-Tab/JSON*	Descrição estruturada de experimentos multiômicos	F ✓ A ✓ I ✓ R ✓	Ontologias *ELIXIR-OBO*; exporta RDF	Curva de aprendizado para ISA-Tab	Implantação em hubs ELIXIR em 2024 X (formerly Twitter)
FAIRDOMHub / SEEK	Gerenciamento e versionamento de projetos	*Repositório web + API REST*	RDM de dados, modelos e protocolos	F ✓ A ✓ I ✓ R ✓	Controle de acesso granular; DOI automático	Requer servidor dedicado ou nuvem	+95 k ativos em 2025 fairsharing.org
FAIR Data Point (FDP)	Publicação semântica de metadados	Catálogo *RDF + SPARQL*	Disponibiliza metadados como *Linked Data*	F ✓ A ✓ I ✓ R ✓	Interoperável com EDC e catálogos CKAN	Exige infraestrutura triplestore	Usado na rede Dutch X-omics 2024 research.rug.nl
FAIR Data Cube (FDCube)	Plataforma integrada *"package"*	*FAIR Data Station + FDP + GraphDB + Vantage6*	*Pipeline* completo *FAIR-at-source* até análise federada	F ✓ A ✓ I ✓ R ✓	Implantação *"in-a-box"*; suporte *Kubernetes*	Configuração inicial complexa	Projeto Trusted World of Corona 2024 BioMed Central
Vantage6	Execução federada com privacidade	*Docker-based computation nodes*	Leva algoritmo até o dado sensível	F ✓ A ✓ I ✓ R -	Mantém dados locais; logs rastreáveis	Depende de contêineres bem definidos	Média > 400 nós ativos em 2025 BioMed Central
Galaxy	Análise bioinformática interativa	*GUI web + Workflows*	QC e análise multiômica sem código	F ✓ A ✓ I ✓ R ✓	*Interface* intuitiva; partilha de históricos	Escala limitada sem cluster SLURM/K8s	Versão 23.2 com suporte FAIR workflows, 2024 Oxford Academic
Nextflow + nf-core	Orquestração de pipelines escaláveis	*DSL2, containers, AWS/SLURM*	*Pipelines* reprodutíveis em nuvem/HPC	F ✓ A ✓ I ✓ R ✓	Execução portátil; >80 *pipelines nf-core*	Requer linha de comando	Hackathon nf-core mar 2025 nf-co.re
OmicsDI	Integração & busca de *datasets*	*ElasticSearch + GraphQL*	Descoberta *cross-repo (GEO, PRIDE…)*	F ✓ A ✓ I ✓ R ✓	Indexa >60 repositórios; API aberta	Falta curadoria manual dos metadados	19 977 datasets indexados (jan 2025) OmicsDI
BioStudies (EMBL-EBI)	Repositório multi-formato	*JSON metadata store*	Armazena estudos heterogêneos	F ✓ A ✓ I - R ✓	Aceita dados "órfãos"; DOI opcional	Pesquisa limitada a texto livre	**

Fonte: elaborado pelos autores (2025).

5.1 Gestão de metadados FAIR

A *suite ISA-Tools*, baseada no modelo *ISA-Tab*, permite descrever estudos experimentais com metadados estruturados e reutilizáveis, promovendo padronização e interoperabilidade (Sansone *et al*., 2021). Já o *FAIRDOMHub* oferece uma infraestrutura de repositório para projetos em ciências da vida, permitindo versionamento e *linkagem* entre dados, modelos e protocolos (Wotstencroft, 2018).

Essas ferramentas utilizam identificadores persistentes (PIDs), como *DOIs* e *ORCIDs*, garantindo rastreabilidade dos dados científicos, conforme recomenda o *DataCite* (Base; Sens; Riedel, 2009).

5.2 Plataformas de análise de dados multiômicos

O *Galaxy* é uma das plataformas mais populares para análise bioinformática, por sua interface amigável e suporte a workflows complexos sem necessidade de programação. É compatível com a maioria dos formatos de dados ômicos e permite compartilhamento de pipelines reprodutíveis (Afgon *et al*., 2018).

O *Nextflow*, por sua vez, oferece escalabilidade e integração com ambientes de nuvem e *containers (Docker, Singularity*), sendo amplamente adotado para análises reprodutíveis de larga escala (Di Tommaso, 2017). Sua integração com *nf-core*, um repositório de pipelines curados, favorece boas práticas FAIR (Ewels *et al*., 2020).

5.3 Integração e compartilhamento

O *OmicsDI (Omics Discovery Index)* é um repositório integrador que permite busca unificada em bancos de dados multiômicos públicos, com ênfase em localização e acesso dos dados (Perez-Riverol, 2017). Ele se conecta a recursos *como PRIDE, GEO e MetaboLights*.

Ferramentas como BioStudies, do EMBL-EBI, também têm ganhado relevância ao reunir dados experimentais heterogêneos sob um mesmo identificador, fortalecendo o reúso e a reprodutibilidade (Sarkans *et al.*, 2021).

5.4 Análises

A adoção de ferramentas compatíveis com FAIR melhora significativamente a eficiência, integridade e impacto dos projetos científicos, o que é fundamental para apoiar as práticas de ciência aberta. Ferramentas como Galaxy são ideais para grupos sem experiência

em programação, enquanto *Nextflow* é recomendado para grupos que desejam maior controle e escalabilidade.

Já para a gestão de metadados, o uso do modelo ISA combinado com *FAIRDOMHub* favorece o cumprimento de requisitos de agências financiadoras e periódicos internacionais. A articulação entre essas plataformas contribui para a construção de ecossistemas digitais que suportam a ciência aberta e colaborativa.

A comparação entre as principais ferramentas e infraestruturas de código aberto voltadas à gestão FAIR de metadados e à análise de dados multiômicos evidencia que a abertura do software é fator decisivo para o avanço de ecossistemas científicos transparentes, colaborativos e reprodutíveis, isto é, alinhados com os valores e princípios da ciência aberta. *Suítes como ISA-Tools, Galaxy, Nextflow + nf-core, FAIR Data Cube e Vantage6 d*emonstram que licenças abertas estimulam inovação comunitária, reduzem barreiras de custo e permitem auditoria pública de código, condição essencial à confiabilidade de pipelines biomédicos sensíveis. Este nível de transparência dialoga com as práticas de uma ciência aberta e reprodutível.

Do ponto de vista de adesão aos princípios FAIR, os frameworks avaliados convergem no uso de identificadores persistentes, modelos de metadados padronizados *(ISA-Tab/JSON, RDF*) e catálogos semânticos (*FAIR Data Point*), tornando dados localizáveis, acessíveis, interoperáveis e reutilizáveis desde a origem. Contudo, a análise revelou que a interoperabilidade plena ainda depende de curadoria ontológica continuada e de investimentos em *triple stores* robustos, áreas onde iniciativas abertas como *OBO Foundry* e *GraphDB Community Edition* têm papel estratégico.

A comparação também mostrou que abordagens federadas, exemplificadas pelo FAIR Data Cube e pelo Vantage6 agregam forte valor em cenários de dados sensíveis, pois mantêm as amostras sob custódia local e enviam algoritmos conteinerizados até os repositórios, preservando privacidade sem sacrificar escalabilidade. Essa arquitetura, aliada à portabilidade de *containers (Docker/Singularity*)e orquestração *Kubernetes*, demonstra que abertura de código + padrões FAIR é um binômio viável para cumprir simultaneamente requisitos de segurança, governança de dados e reprodutibilidade científica.

Entretanto, persistem desafios críticos: (i) curva de aprendizagem para modelagem de metadados ISA e configuração de nós federados; (ii) lacunas de infraestrutura em nuvens acadêmicas de baixa renda; e (iii) necessidade de programas de capacitação que integrem

bioinformática, ciências de dados e gestão FAIR. Políticas institucionais de financiamento contínuo e métricas de avaliação que reconheçam o compartilhamento aberto de pipelines e dados tendem a acelerar a adoção dessas soluções.

Esses resultados dialogam diretamente com o movimento da Ciência Aberta, especialmente na Ciência da Informação, ao enfatizar a importância da interoperabilidade, do acesso aberto e da gestão transparente de dados. A adoção de princípios FAIR e o uso de ferramentas abertas fortalecem a representação e a organização do conhecimento, pilares teórico-metodológicos centrais à área.

6 CONSIDERAÇÕES FINAIS

A adoção de ferramentas digitais alinhadas aos princípios FAIR e alinhados aos princípios da ciência aberta, representam os pilares para o avanço científico, especialmente na saúde, onde a complexidade e a sensibilidade dos dados exigem infraestruturas robustas, éticas e interoperáveis. Ao garantir que dados multiômicos sejam estruturados desde a origem com metadados ricos, identificadores persistentes e protocolos abertos, essas ferramentas não apenas ampliam a transparência e a reprodutibilidade das pesquisas, mas também democratizam o acesso ao conhecimento, acelerando descobertas em medicina personalizada, diagnóstico precoce e estratégias de saúde pública.

A gestão de dados multiômicos, alinhada aos princípios FAIR, é um imperativo estratégico para o avanço científico contemporâneo. A análise comparativa realizada neste estudo demonstra que a adoção de ferramentas e infraestruturas abertas, como *Galaxy, Nextflow, FAIRDOMHub, ISA-Tools e OmicsDI*, não apenas atende a critérios técnicos rigorosos, mas também viabiliza a estruturação de fluxos de trabalho robustos e reprodutíveis. Essas soluções garantem que dados complexos sejam localizáveis, acessíveis e semanticamente interoperáveis desde sua origem, reduzindo lacunas entre a geração e o reúso de informações em larga escala, um avanço crítico para áreas como a genômica e a saúde digital.

A integração entre plataformas FAIR amplia seu potencial, criando ecossistemas digitais seguros e adaptáveis. Por exemplo, a combinação de captura de metadados *FAIR-at-source (via ISA-Tools), pipelines* conteinerizado*s (Nextflow*) e execução federada (*FAIR Data Cube*) permite análises distribuídas sem comprometer a privacidade ou a integridade dos dados sensíveis. Essa sinergia não só otimiza a reprodutibilidade e reduz redundâncias, como

também acelera descobertas colaborativas, especialmente em projetos multicêntricos que demandam interoperabilidade semântica e governança compartilhada.

No campo da saúde, a implementação dessas ferramentas assume relevância transformadora. A rastreabilidade analítica, a integridade dos dados, a interoperabilidade e a capacidade de reutilização de conjuntos multiômicos são fundamentais para avanços em medicina personalizada, diagnóstico precoce de doenças e políticas públicas baseadas em evidências. Plataformas como *OmicsDI*, que integram repositórios heterogêneos, ou o *FAIR Data Cube*, que viabiliza análises federadas, exemplificam como a gestão FAIR pode traduzir *Big Data* em *insights* clínicos e terapêuticos, beneficiando diretamente a prática médica e a saúde populacional.

Contudo, a plena realização desse potencial exige superar desafios estruturais, tais como a necessidade de capacitação de pesquisadores em boas práticas FAIR, os custos elevados de armazenamento de dados em larga escala e a heterogeneidade nos formatos e ontologias adotadas por diferentes repositórios. Para superá-los, investimentos contínuos em capacitação técnica são essenciais para democratizar ferramentas complexas, como modelagem de metadados e configuração de ambientes federados, enquanto políticas institucionais devem incentivar a adoção de padrões FAIR em todas as etapas da pesquisa, do desenho experimental à publicação de dados. Além disso, a integração tecnológica em infraestruturas de nuvem acessíveis, aliada à harmonização de ontologias, é fundamental para reduzir desigualdades no acesso a recursos computacionais avançados. Essas medidas, articuladas, são críticas para consolidar um ecossistema de pesquisa robusto, equitativo e alinhado com as demandas científicas contemporâneas.

Em síntese, a transição para uma ciência orientada por dados FAIR não é apenas uma evolução técnica, mas um compromisso ético com a eficiência, a transparência e a equidade global. Ao priorizar ferramentas abertas e infraestruturas interoperáveis, a comunidade científica pode transformar o imenso volume de dados multiômicos em conhecimento aplicável, acelerando inovações que impactam diretamente a vida das pessoas. Nas ciências da vida e da saúde, essa abordagem representa não apenas um diferencial estratégico, mas uma responsabilidade coletiva em garantir que a revolução dos dados beneficie a sociedade de forma inclusiva e sustentável.

As pesquisas com dados ômicos, corretamente fundamentadas nos princípios da ciência aberta e alinhadas aos princípios FAIR, utilizam identificadores que garantem a

localização, acessibilidade, interoperabilidade e reutilização dos dados. Essa abordagem confere credibilidade às pesquisas e assegura citações e crédito adequados aos pesquisadores envolvidos. Nesse processo, a preservação e segurança dos dados sensíveis agrega valor significativo aos cenários de investigação. Contudo, além de garantir a proteção desses dados sensíveis, é fundamental implementar mecanismos rigorosos de controle de acesso, elementos fundamentais para a consolidação de uma ciência aberta.

Essas pesquisas – que abrangem áreas como genômica, transcriptômica, proteômica, metabolômica e epigenômica – oferecem possibilidades que vão desde a pesquisa básica até aplicações clínicas e industriais. Por meio delas, é possível realizar investigações biomédicas para desvendar mecanismos de doenças complexas (como câncer, *Alzheimer* ou diabetes), identificar biomarcadores, desenvolver medicina personalizada e impulsionar a criação de novos fármacos. Constituem, portanto, um elemento fundamental para novas perspectivas no desenvolvimento de tratamentos personalizados e novos fármacos objetivando atender as necessidades de saúde da população mundial.

REFERÊNCIAS

AZEVEDO, N. H.; MENDONÇA, P. C. C. Dados abertos na pesquisa em educação em ciências: perspectivas, desafios e possibilidades. **Ensaio Pesquisa em Educação em Ciências**, Belo Horizonte, v. 26, p. e51515, 2024.

BRASE, J.; SENS, I.; RIEDEL T. The metadata schema for the publication and citation of research data. **Data Sci J**, v. 8, p.1–16, 2009.

DEIST, Thomas M. *et al*. Distributed learning on 20 000+ lung cancer patients – The Personal Health Train. **Radiotherapy and Oncology**, [*s. l.*], v. 144, p. 189–200, 2020. Disponível em: DOI: https://doi.org/10.1016/j.radonc.2019.11.019. Acesso em: 11 jul. 2025.

DI TOMMASO, Paolo *et al.* Nextflow enables reproducible computational workflows. **Nature Biotechnology**, [*s. l.*], v. 35, n. 4, p. 316–319, 2017. Disponível em: DOI: https://doi.org/10.1038/nbt.3820. Acesso em: 11 jul. 2025.

EWELS, Philip A. *et al.* The nf-core framework for community-curated bioinformatics pipelines. **Nature Biotechnology**, [*s. l.*], v. 38, p. 276–278, 2020. Disponível em: DOI: https://doi.org/10.1038/s41587-020-0439-x. Acesso em: 11 jul. 2025.

HADDAWAY, N. R. Síntese aberta: sobre a necessidade de síntese de evidências para adotar a ciência aberta. **Environmental Evidence**, v. 7, p. 26, 2018. DOI: https://doi.org/10.1186/s13750-018-0140-4. Disponível em: https://environmentalevidencejournal.biomedcentral.com/articles/10.1186/s13750-018-0140-4. Acesso em: 11 jul. 2025.

HASIN, Yehuda; SAGI, Dan; MARIAN, Avner. Multi-omics approaches to disease. **Genome Biology**, [*s. l.*], v. 18, n. 1, p. 83, 2017. Disponível em: DOI: https://doi.org/10.1186/s13059-017-1215-1. Acesso em: 11 jul. 2025.

JACOBSEN, Anne *et al*. A Generic Workflow for the Data FAIRification Process. **Data Intelligence**, [*s. l.*], v. 2, n. 1–2, p. 56–65, 2020. Disponível em: DOI: https://doi.org/10.1162/dint_a_00028. Acesso em: 11 jul. 2025.

MONS, Barend *et al.* Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. **Information Services & Use**, [*s. l.*], v. 40, n. 1–2, p. 1–8, 2020. Disponível em: DOI: https://doi.org/10.3233/ISU-200084. Acesso em: 11 jul. 2025.

PEREZ-RIVEROL, Yasset *et al.* Discovering and linking public omics data sets using the Omics Discovery Index. **Nature Biotechnology**, [*s. l.*], v. 35, n. 5, p. 406–409, 2017. Disponível em: DOI: https://doi.org/10.1038/nbt.3790. Acesso em: 11 jul. 2025.

SANSONE, Susanna-Assunta *et al*. FAIRsharing as a community approach to standards, repositories and policies. **Nature Biotechnology**, [*s. l.*], v. 37, n. 4, p. 358–367, 2019. Disponível em: DOI: https://doi.org/10.1038/s41587-019-0080-8. Acesso em: 11 jul. 2025.

SANSONE, Susanna-Assunta *et al.* FAIR principles: interpretations and implementation considerations. **Data Intelligence**, [*s. l.*], v. 3, n. 1, p. 10–29, 2021. Disponível em: DOI: https://doi.org/10.1162/dint_r_00024. Acesso em: 11 jul. 2025.

SARKANS, Ugis *et al*. BioStudies: an updated resource for curating and sharing multi-omics data. **Nucleic Acids Research**, [*s. l.*], v. 49, n. D1, p. D1225–D1231, 2021. Disponível em: DOI: https://doi.org/10.1093/nar/gkaa993. Acesso em: 11 jul. 2025.

STEPHENS, Zachary D. *et al*. Big Data: Astronomical or Genomical? **PLoS Biology**, [*s. l.*], v. 13, n. 7, e1002195, 2015. Disponível em: DOI: https://doi.org/10.1371/journal.pbio.1002195. Acesso em: 11 jul. 2025.

UNESCO. **Recomendação sobre ciência aberta.** Paris: Organização das Nações Unidas para a Educação, a Ciência e a Cultura, 2022. Disponível em: https://unesdoc.unesco.org/ark:/48223/pf0000379949_por . Acesso em: 11 jul. 2025.

VAN VLIET, Kim; MOORE, Claybourne. Citizen science initiatives: engaging the public and demystifying science. **Journal of Microbiology & Biology Education**, v. 17, n. 1, p. 13–16, 2016. DOI: 10.1128/jmbe.v17i1.1019. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4798796/. Acesso em: 11 jul. 2025.

WILKINSON, Mark D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. **Scientific Data**, [*s. l.*], v. 3, p. 160018, 2016. Disponível em: DOI: https://doi.org/10.1038/sdata.2016.18. Acesso em: 11 jul. 2025.

WITJES, Sander *et al*. FAIR Data Cube: An Architecture for FAIR Federated Multi-Omics Data Analysis. *In*: INTERNATIONAL CONFERENCE ON BIOMEDICAL AND HEALTH INFORMATICS (ICBHI), 5., 2022, Concepción, Chile: IEEE. **Proceedings** […], Concepción, Chile: [*s. n*.], 2022. p. 1–4. Disponível em: DOI: https://doi.org/10.1109/ICBHI57336.2022.10030310. Acesso em: 11 jul. 2025.