# Rafael Joia

**Master in Software Engineering**
**Towards Reproducible and Privacy-preserving Analyses Across Federated Repositories for Omics data**
**Rafael Lima Joia**
—

## Abstract

Even when duly anonymized, health research data has the potential to be disclosive and therefore requires special safeguards according to the European General Data Protection Regulation (GDPR). Furthermore, the incorporation of FAIR principles (Findable, Accessible, Interoperable, Reusable) for a more favorable reuse of existing data, calls for an approach where sensitive data is kept locally and only metadata and aggregated results are shared. Additionally, since central pooling is discouraged by ethical, legal, and societal issues, it is more frequent to observe maturing data management frameworks, and platforms adopting the federated approach.

Current implementations of privacy-preserving analysis frameworks seem to be limited when data becomes very large (millions of rows, hundreds of variables). Biological samples data, collected by high-throughput technologies, such as Next Generation Sequencing (NGS), and processed by computational workflows known as bioinformatics pipelines, are examples of this kind of data (commonly known as Omics data). The reproducibility of these pipelines is hard and it is often underestimated. Nevertheless, it is important to generate trust in scientific results, and therefore, is fundamental to know how these Omics data were generated or obtained.

This work will leverage the promising results of current open-source implementations for distributed privacy-preserving analyses, while aiming at generalizing the approach and addressing some of their shortcomings, including the reproducibility concerns.

The results were promising, seeing that the privacy-preserving analysis was effective when using the DataSHIELD framework in conjunction with the "resource R" package. We also concluded that the adoption of specialized DataSHIELD packages for Omics analyses, such as dsOmics, is a viable pathway to leverage the privacy-preserving for Omics data. To address the reproducibility challenges, we defined a relational database model to represent the steps, commands and operations executed by the bioinformatics pipelines. The proposed reproducible relation model can afford the traceability of bioinformatics pipelines very well, but this model alone does not guarantee a full reproducible ecosystem, since it does not solve the platform isolation problem. It can only be guaranteed when combining reproducible tools that offer built-in support for containers, such as Nextflow or Snakemake, and a set of values and good practices.

We concluded that the proposed solution would be a viable option for privacy-preserving analysis using Omics data. In contrast, the proposed pipeline reproducibility model must be improved or incorporated into existing reproducible pipeline tools.

## Resumo

Os dados de pesquisas em saúde, mesmo quando devidamente anonimizados, têm o potencial de ser reveladores e, portanto, requerem salvaguardas especiais de acordo com o Regulamento Geral Europeu de Proteção de Dados (GDPR). Por outro lado, a incorporação dos princípios FAIR (Findable, Accessible, Interoperable, Reusable) para a reutilização mais favorável dos dados existentes exige uma abordagem em que os dados privados sejam mantidos localmente e apenas metadados e resultados agregados sejam compartilhados. Adicionalmente, como o agrupamento central de dados é desencorajado por questões éticas,

legais e sociais, é mais frequente observar frameworks de gerenciamento de dados e plataformas adotando uma abordagem federada.

As implementações atuais de *frameworks* de análise de preservação de privacidade parecem ser limitadas quando o volume de dados se torna muito grande (milhões de linhas, centenas de variáveis). Dados de amostras biológicas, coletadas por tecnologias de alto desempenho, como as Next Generation Sequence (NGS), e processadas por *workflows* computacionais conhecidos como pipelines bioinformáticos, são exemplos deste tipo de dado (comummente conhecidos como dados ômicos). A reprodutibilidade desses *pipelines* é difícil e muitas vezes subestimada. No entanto, ela é importante para gerar confiança nos resultados científicos e, portanto, é fundamental saber como esses dados ômicos foram gerados ou obtidos.

Este trabalho aproveitará os resultados promissores das implementações atuais de código aberto para análises distribuídas de preservação de privacidade, ao mesmo tempo que visa generalizálas, abordando algumas de suas deficiências, incluindo preocupações sobre reprodutibilidade.

Os resultados foram promissores, visto que a análise de preservação de privacidade foi eficaz ao usar a estrutura DataSHIELD em conjunto com o pacote "resource R". Também concluímos que a adoção de pacotes DataSHIELD especializados para análises que envolvam dados ômicos é um caminho viável para alavancar a preservação da privacidade deste tipo de dado. Para enfrentar os desafios de reprodutibilidade, definimos um modelo de banco de dados para representar as etapas, comandos e operações executadas pelos *pipelines* bioinformáticos. O modelo relacional proposto pode permitir a rastreabilidade dos *pipelines* muito bem, mas esse modelo sozinho não garante um ecossistema totalmente reprodutível, uma vez que não resolve o problema de isolamento das plataformas. Isto só pode ser garantido ao combinar ferramentas que oferecem suporte integrado para *containers*, como Nextflow ou Snakemake, e um conjunto de valores e boas práticas.

Concluímos que a solução proposta seria uma opção viável para análise de preservação de privacidade usando dados ômicos. Em contraste, o modelo proposto de reprodutibilidade deve ser melhorado ou incorporado às ferramentas existentes de reprodutibilidade de *pipelines*.

## Jury

- Chair: Prof. Nuno Honório Rodrigues Flores,
- External Examiner: Prof. Gur Yaari
- Supervisor: Prof. João Correia Lopes
- Date: 14/07/2021