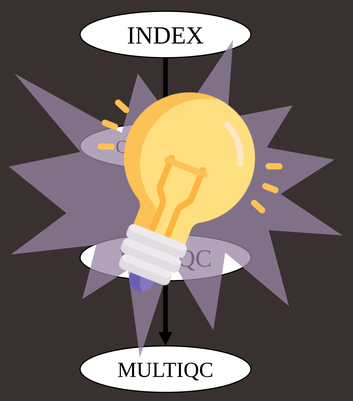


From Headache to High-Five! We are trying to make **reproducibility in Bioinformatics** easier

An initial exploration into **navigating fine-grained provenance** and using its precious information

BOSSUT Noémie - 2024

Background

In the field of bioinformatics, where decisions can profoundly impact lives, workflows need to embody more **FAIR principles**: **F**indability, **A**ccessibility, **I**nteroperability, and **R**eusability.

Workflow managers like Nextflow and Snakemake offer tools to boost FAIRness. However, each process operates as a **black box**, making it challenging to leverage fine-grained provenance despite its potential for **transparency**, **debugging**, and **reusability**.

This study explores the feasibility **of extracting fine-grained provenance from intermediate files**, raising questions about the optimal approaches for its storage and visualization.

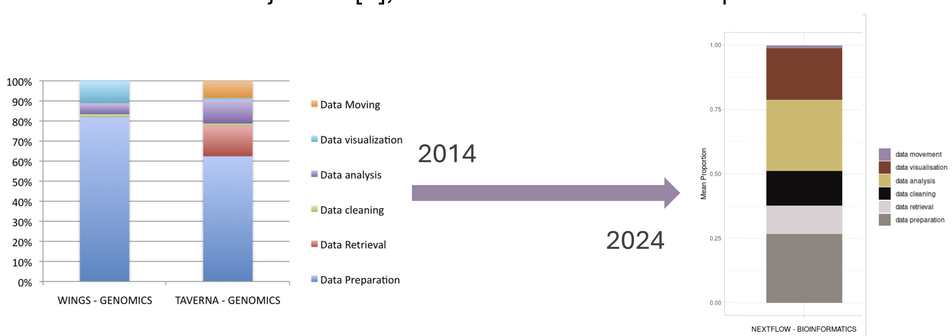
Methods

nextflow

nf-core

8 Nextflow workflows from nf-core were executed (representing **151 processes)**, with each intermediate output manually analyzed to investigate the potential extractability of fine-grained provenance and its mapping type, specifically focusing on text files.

Additionally, for each process, we determined the motifs it relies on, referencing the framework of Garijo *et al*.[1], to observe the evolution of practices.



Results

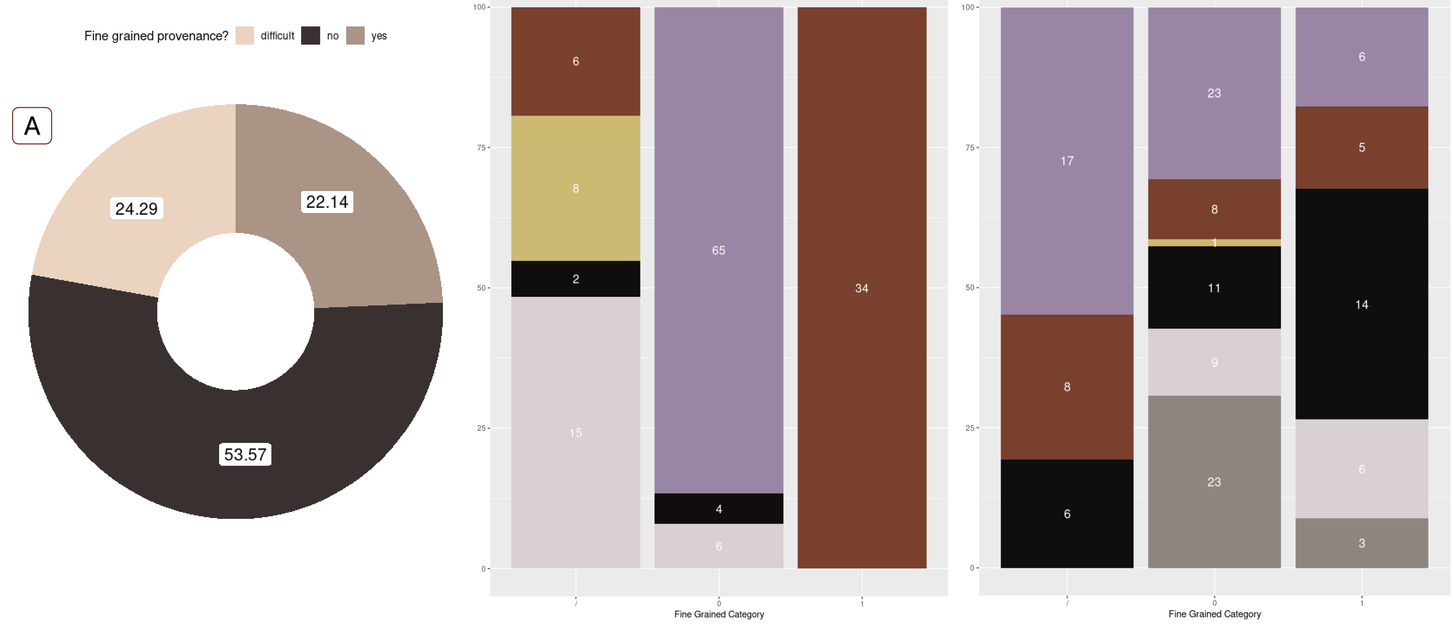


Figure A : Distribution of Processes with Accessible Fine-Grained Provenance

Figure B : Mapping Type Distribution Considering Fine-Grained Provenance Extractability

Figure C : Distribution of Process Motifs Based on Fine-Grained Provenance Accessibility

In **nearly 50% of processes, fine-grained provenance can be inferred** and/or extracted. Simple **one-to-one mapping provenance can consistently be extracted** (albeit with varying difficulty). Feasibility for **fine-grained provenance extraction extends across all types of work**, including visualization

Access to the pdf, additional figures and (hopefully) future publications here!

[1] Garijo, Daniel, Pinar Alper, Khalid Belhajjame, Oscar Corcho, Yolanda Gil, et Carole Goble. « Common Motifs in Scientific Workflows: An Empirical Analysis », 2014.

LAMSADE UMR 7243 Dauphine | PSL UNIVERSITÉ PARIS

ShareFAIR

INSTITUT PASTEUR

FRANCE 2030