# Data Management Practices in Bioinformatic Workflows: A Brief Overview of Current Approaches


Wesley Brashear, Ph.D.

11 July 2025

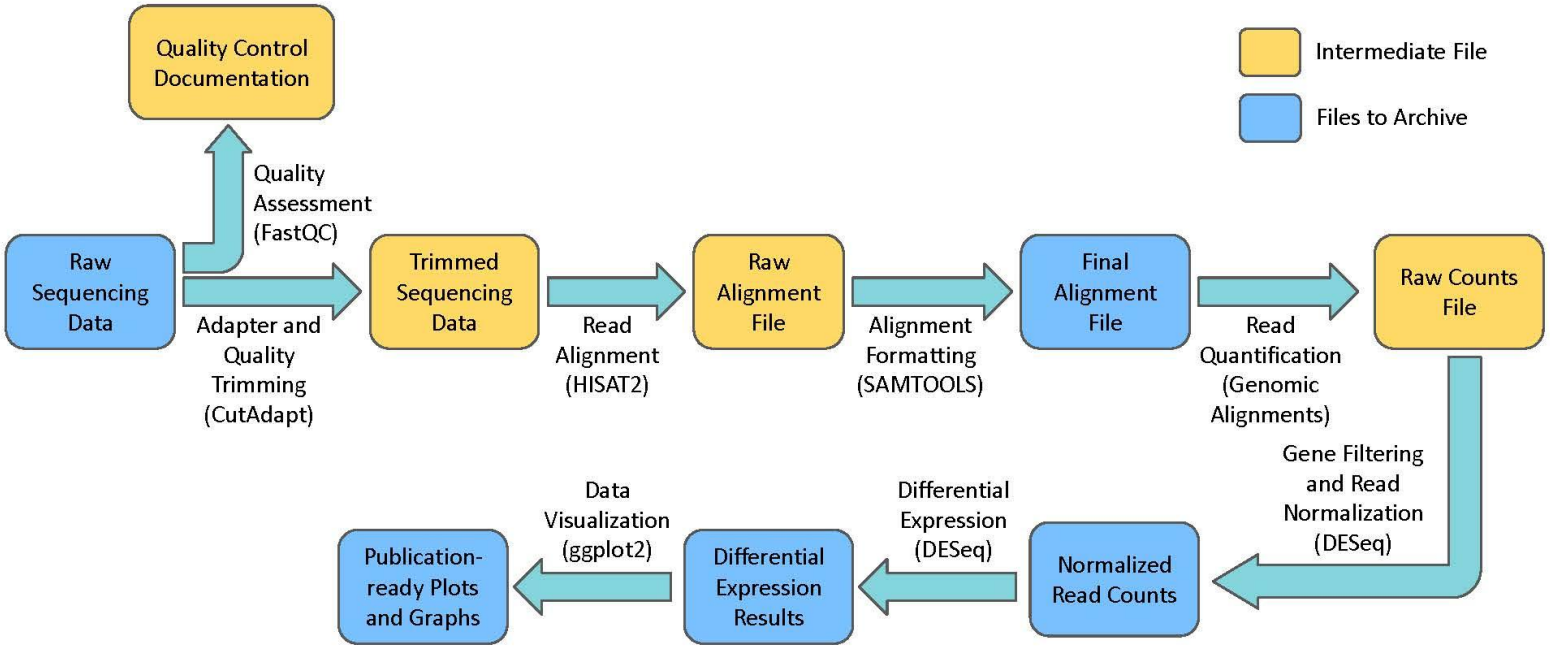
# Introduction

- Reproducibility is a core component of the scientific process
- Well-established practices for recording and reporting experimental details in the life sciences
- Next generation sequencing (NGS) revolutionized biological research
- New computational methods and massive data sets disrupted traditional data provenance and management practices

# Introduction

- Existing data repositories were massively expanded
- Labs quickly established protocols or allowed new research assistants or graduate students to handle RDM independently
- Standard approaches and protocols have emerged and are being rapidly adopted

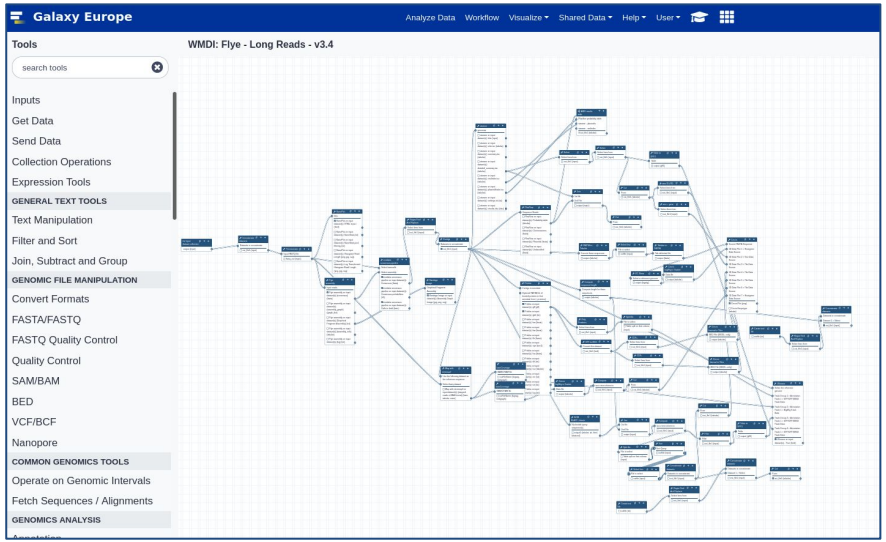

# Introduction



# Current Approaches

## Galaxy

- Graphical User Interface
- Data and analysis history managed and saved automatically
- Workflows can be saved and shared
- Galaxy Tool Shed houses thousands of published workflows
- Some users find it restrictive and cumbersome



# Current Approaches

## Code Repositories

- Many researchers create GitHub repositories for each project
- Deposit job scripts used for each analysis, relevant results, and any relevant information
- Maximum flexibility
- Researcher-dependent reproducibility (not good)
- Issues with repository ownership (e.g. members leaving lab)


# Current Approaches

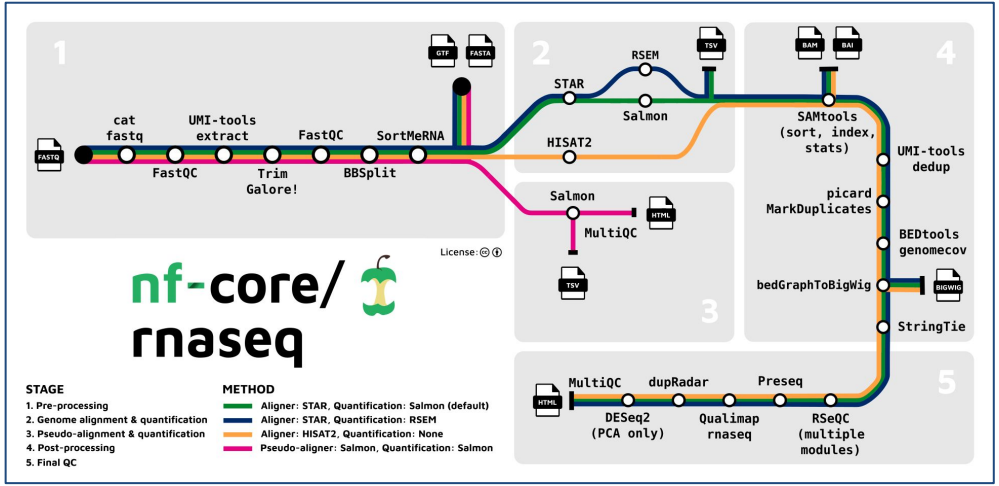
## Snakemake

- Python-based workflow management system that allows users to easily create and share workflows
- Broad user community that is rapidly growing
- Provide access to curated and citable workflows
- Integrates with Jupyter (data visualization, report generation)
- Integrates well with batch scheduling systems

# Current Approaches

## Nextflow

- Java/Groovy-based workflow management system
- Native task support, workflow versioning, container management
- Built-in batch schedulers and distributed clusters
- Integrates with GitHub/Bitbucket
- Provides curated and citable workflows



# Conclusions

- RDM practices and software preferences vary from lab to lab
- Snakemake and Nextflow offer great balance of flexibility and structure
- Increasing adoption of more standardized approaches
- Need to incorporate RDM in informal and formal life sciences/bioinformatics training