

Workflow Management

Epigenomics Data Analysis Workshop

Agata Smialowska 2025

Scientific Data Analysis

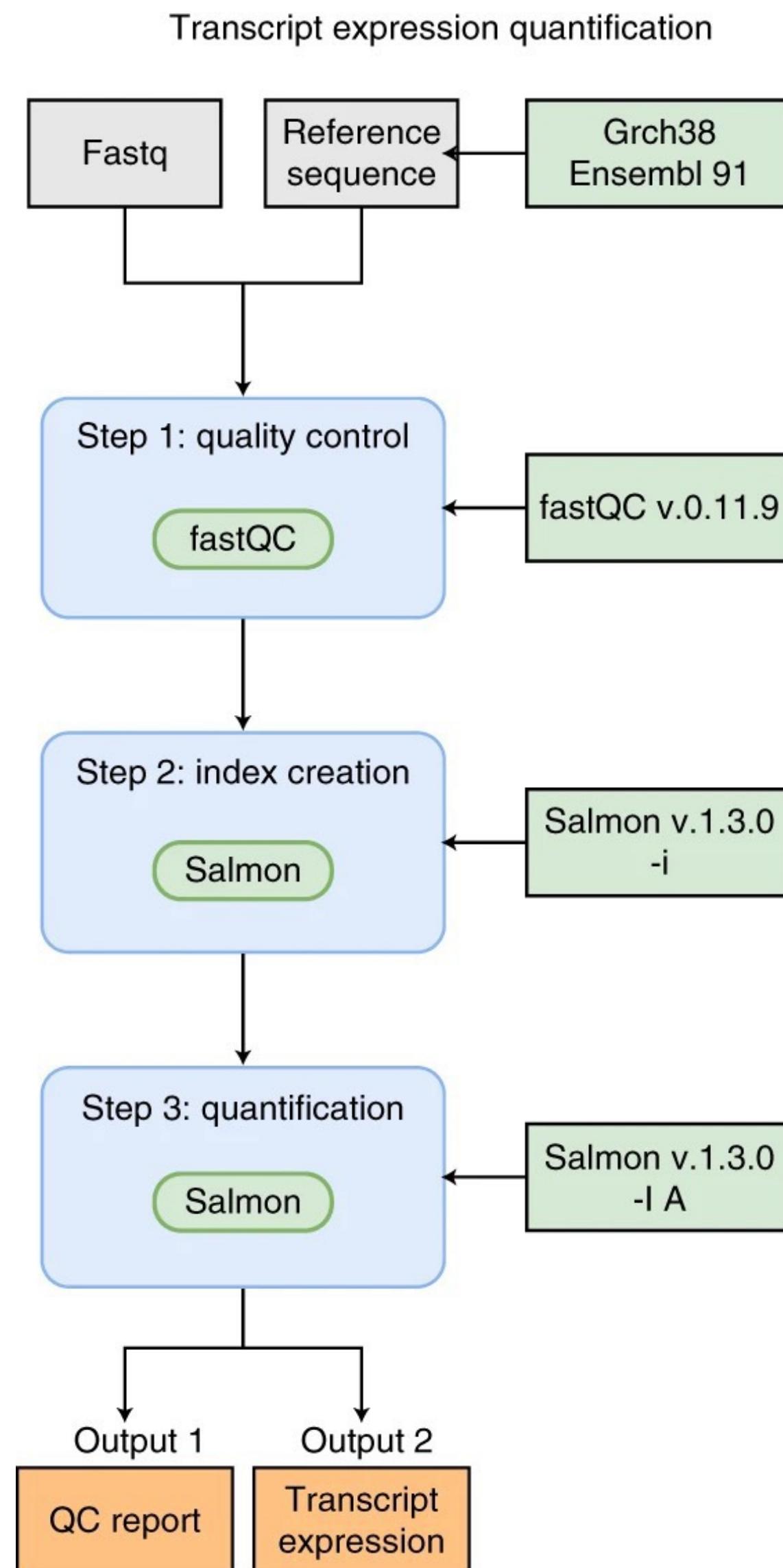
Perspective | Published: 23 September 2021

Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers

[Laura Wratten](#), [Andreas Wilm](#) & [Jonathan Göke](#) 

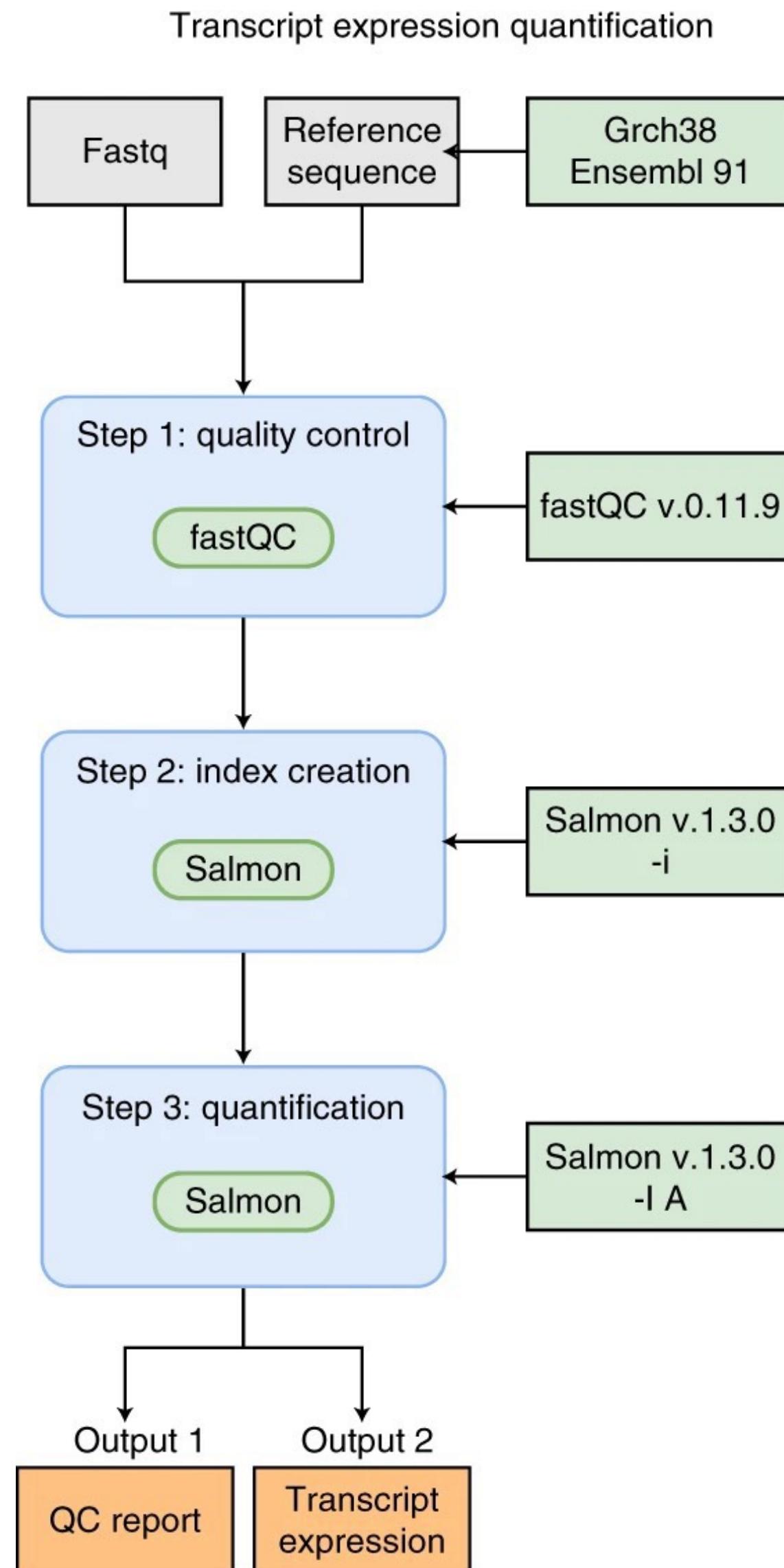
[Nature Methods](#) **18**, 1161–1168 (2021) | [Cite this article](#)

Scientific Data Analysis



- Many tools
 - parameter space
 - versions
 - dependencies
- References, annotations, databases
 - versions
- Resource management
- Manage execution
 - reruns

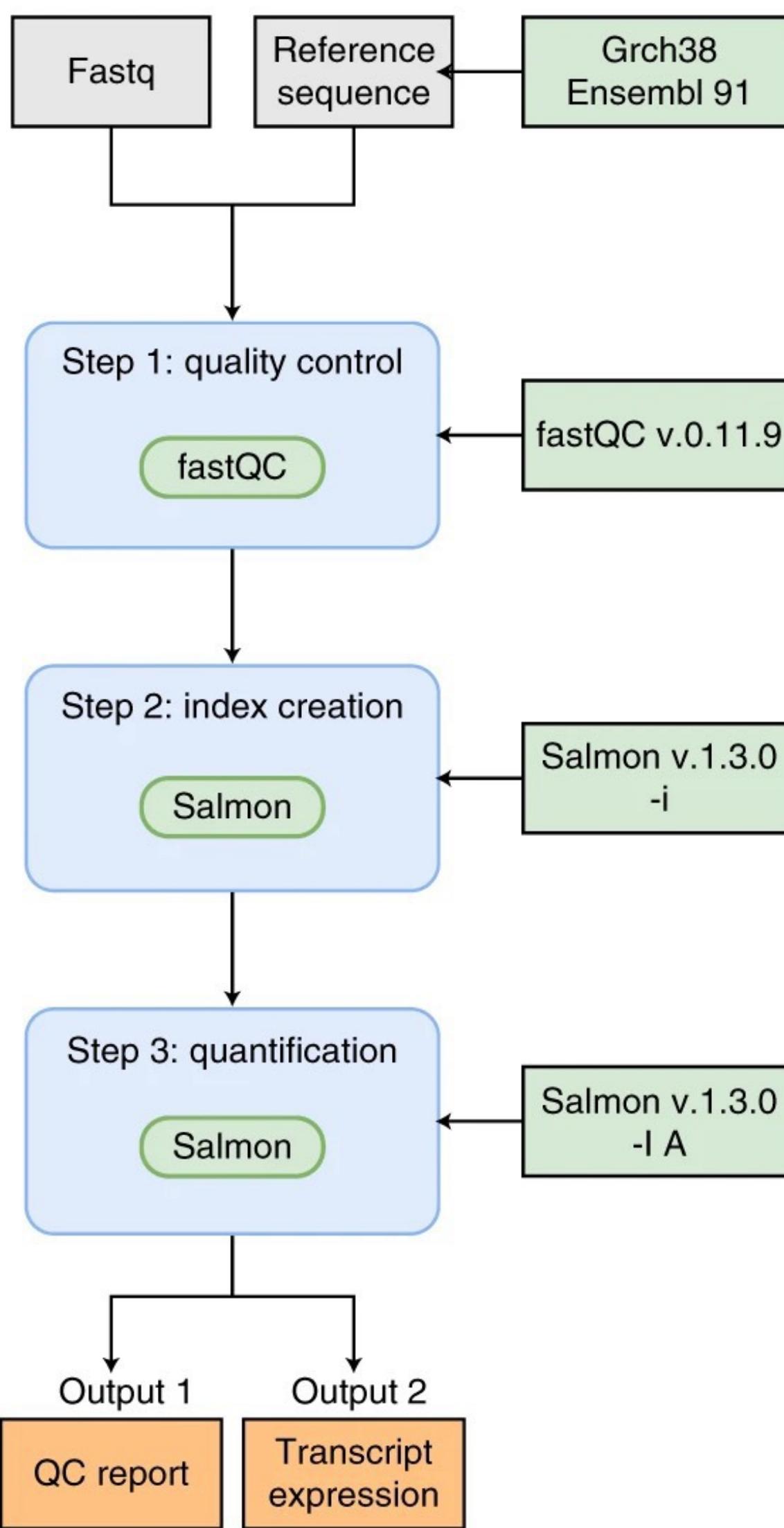
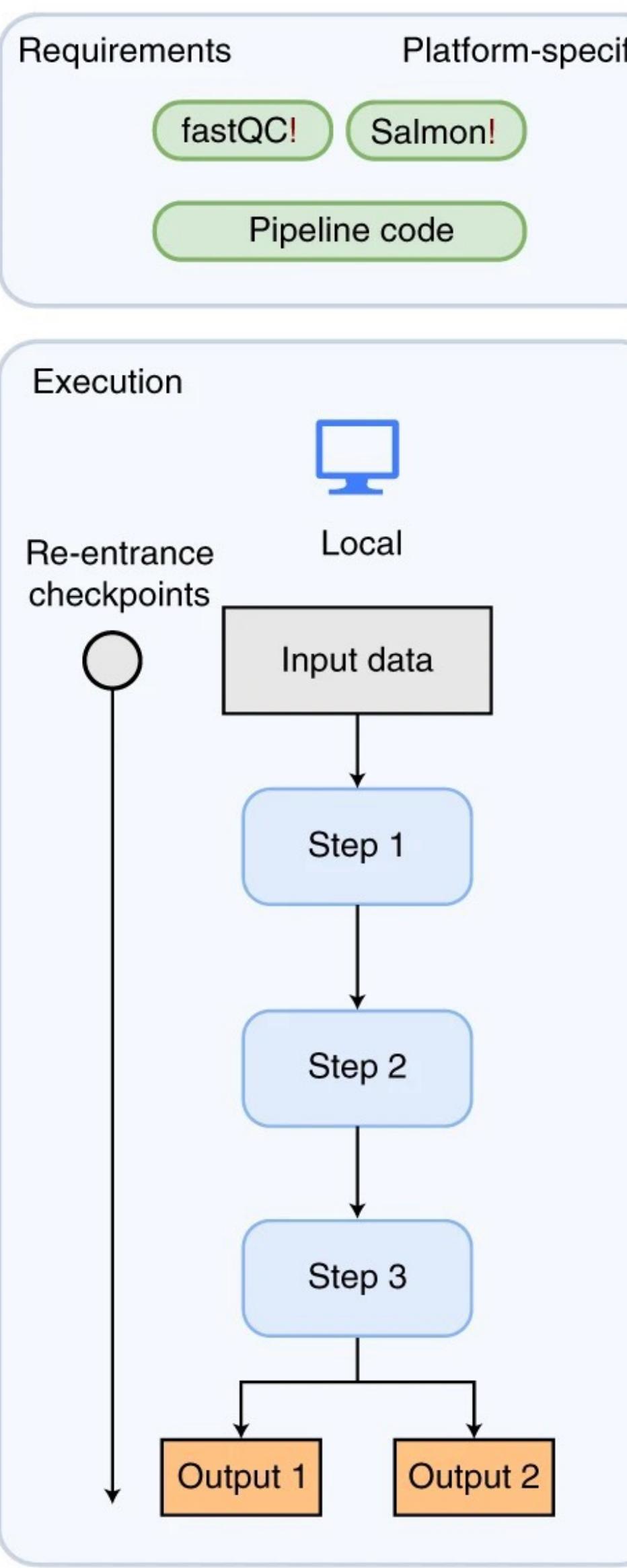
Scientific Data Analysis



- Many tools
 - parameter space
 - versions
 - dependencies
 - Large workflows
 - References, annotations, databases
 - Analysis reruns
 - versions
 - Collaboration
 - Resource management
 - Reproducibility
 - Manage execution
 - reruns
- Not trivial**

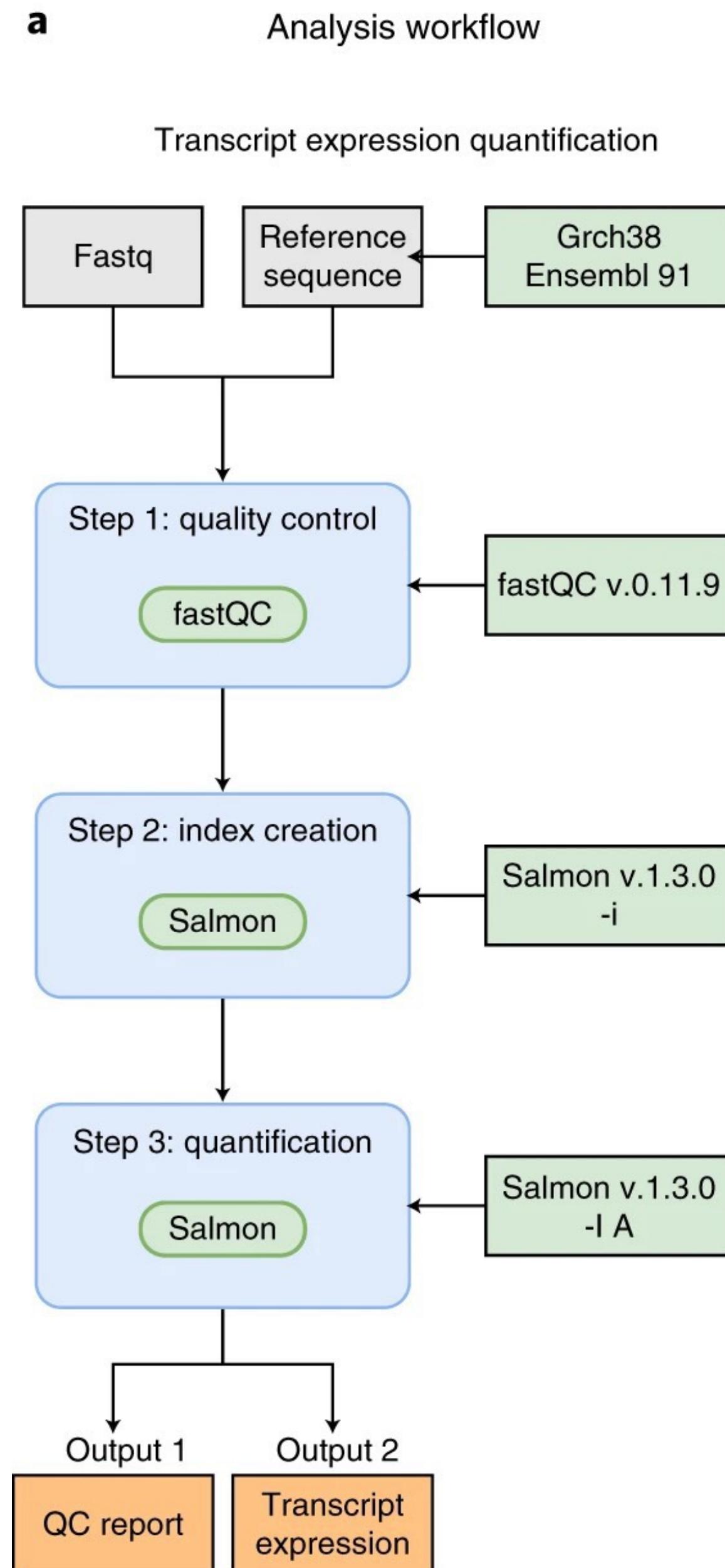
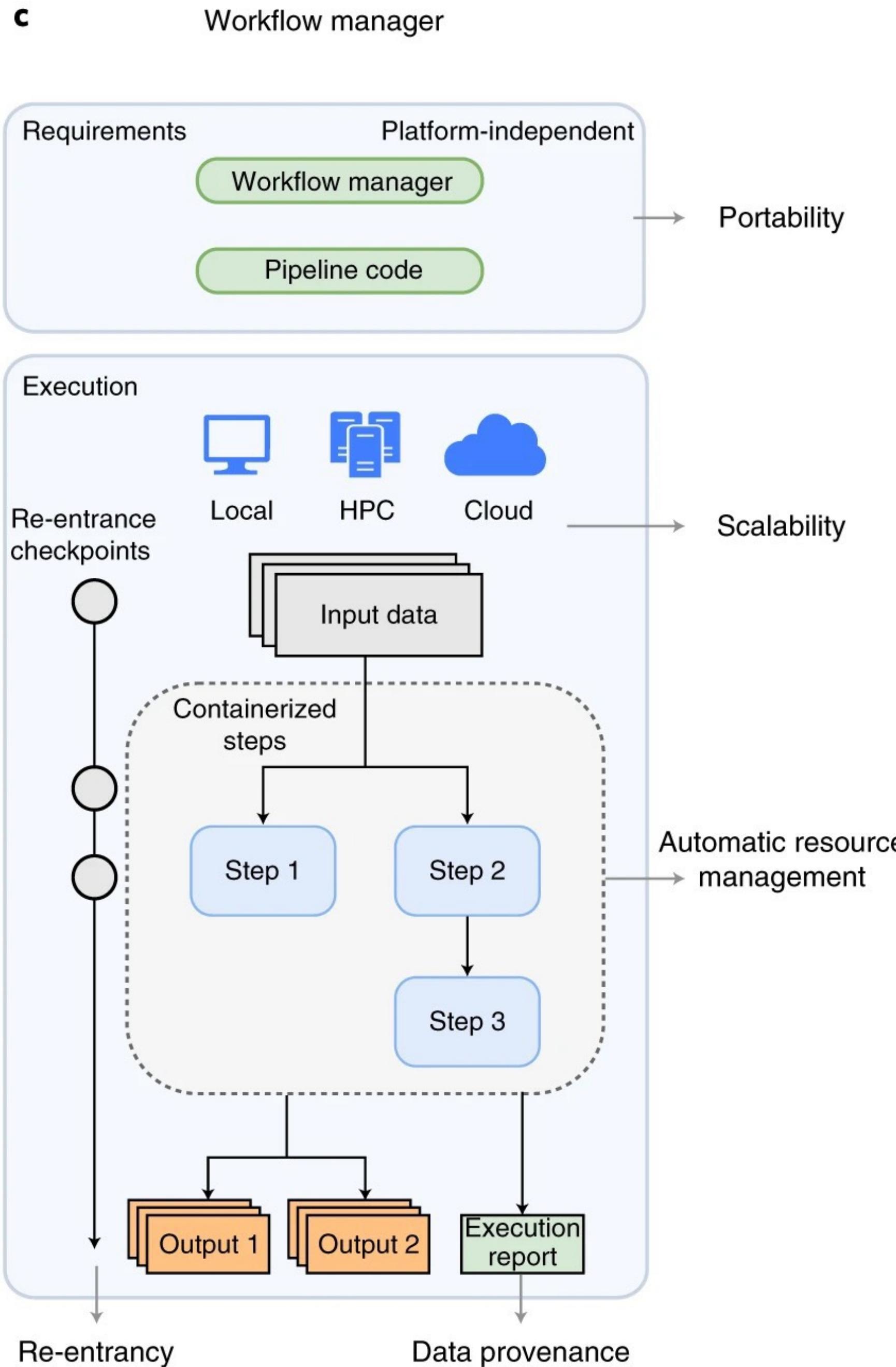
a Analysis workflow

Transcript expression quantification

**b** Traditional pipeline

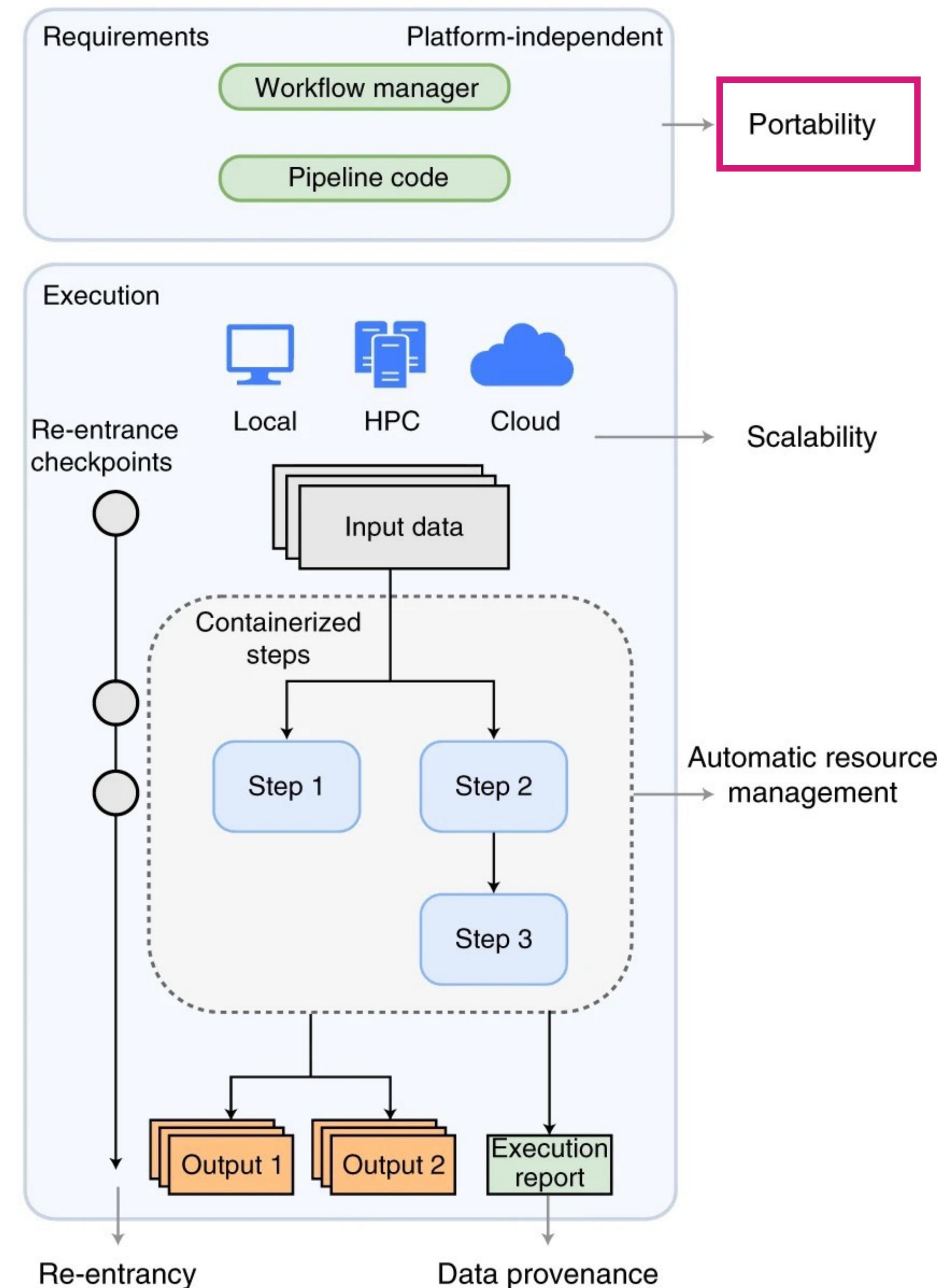
Bash scripts to chain multiple tools to make the analyses easier

- Cannot resume failed run
- Connected to local infrastructure
- Lack of documentation
 - parameters
 - versions
- Low portability
- Cumbersome maintenance
 - tools
- workflow

a**c**

Workflow manager to coordinate input and output of multiple tools to automate the analysis

- Automatic resource management
 - resume failed runs
 - interaction with HPC scheduling system, task resubmission
 - configurable
- Tool management
 - parameters
 - containerised environments
- Fully portable and scalable
- Documentation



Portability and Reproducibility

execution across different platforms and over time

- Package Managers

- automate installing and manage software dependencies

CONDA



- Software containers

- isolated execution environment

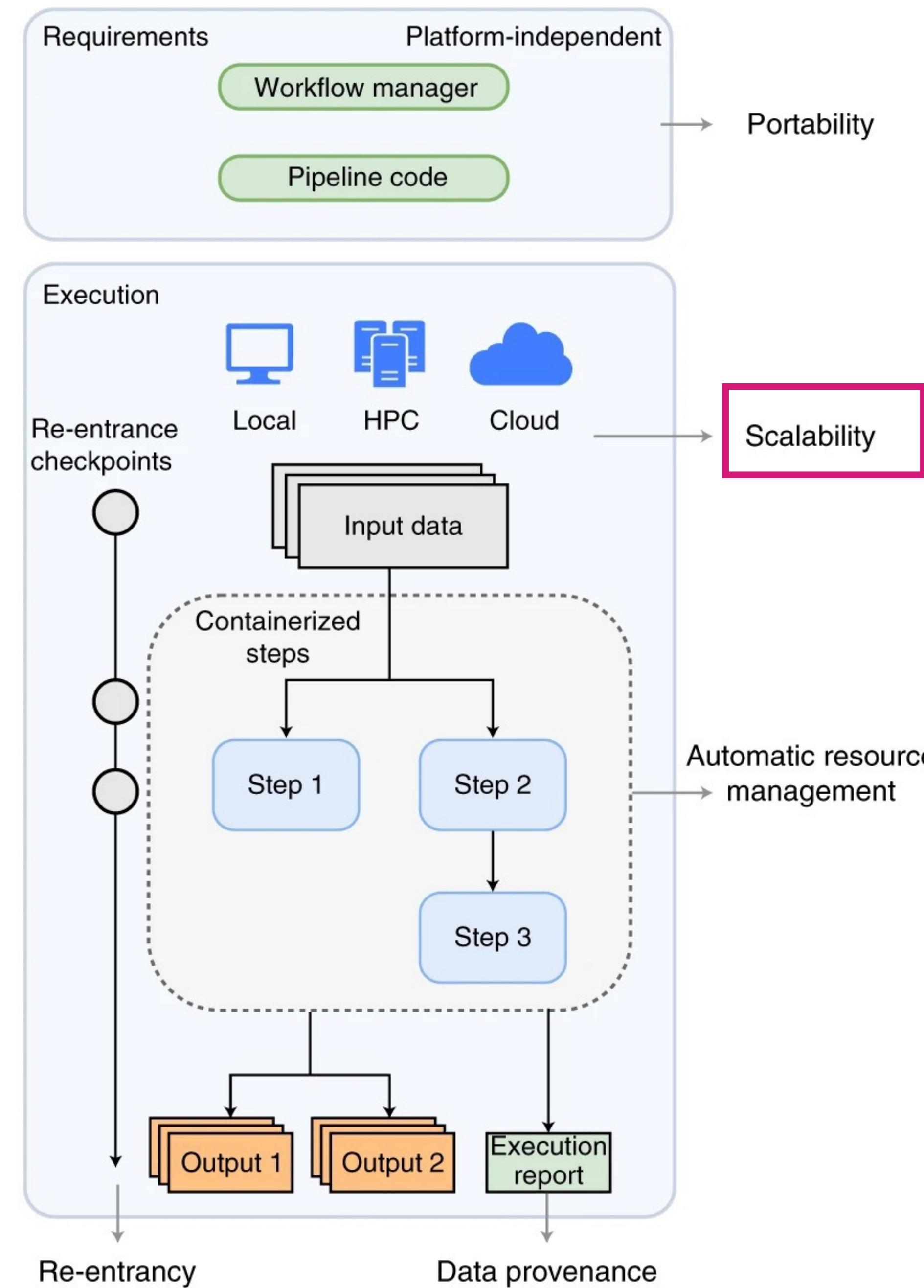
APPTAINER



- BioContainers: prebuilt containers of all major tools

OS and software versions can affect the results.

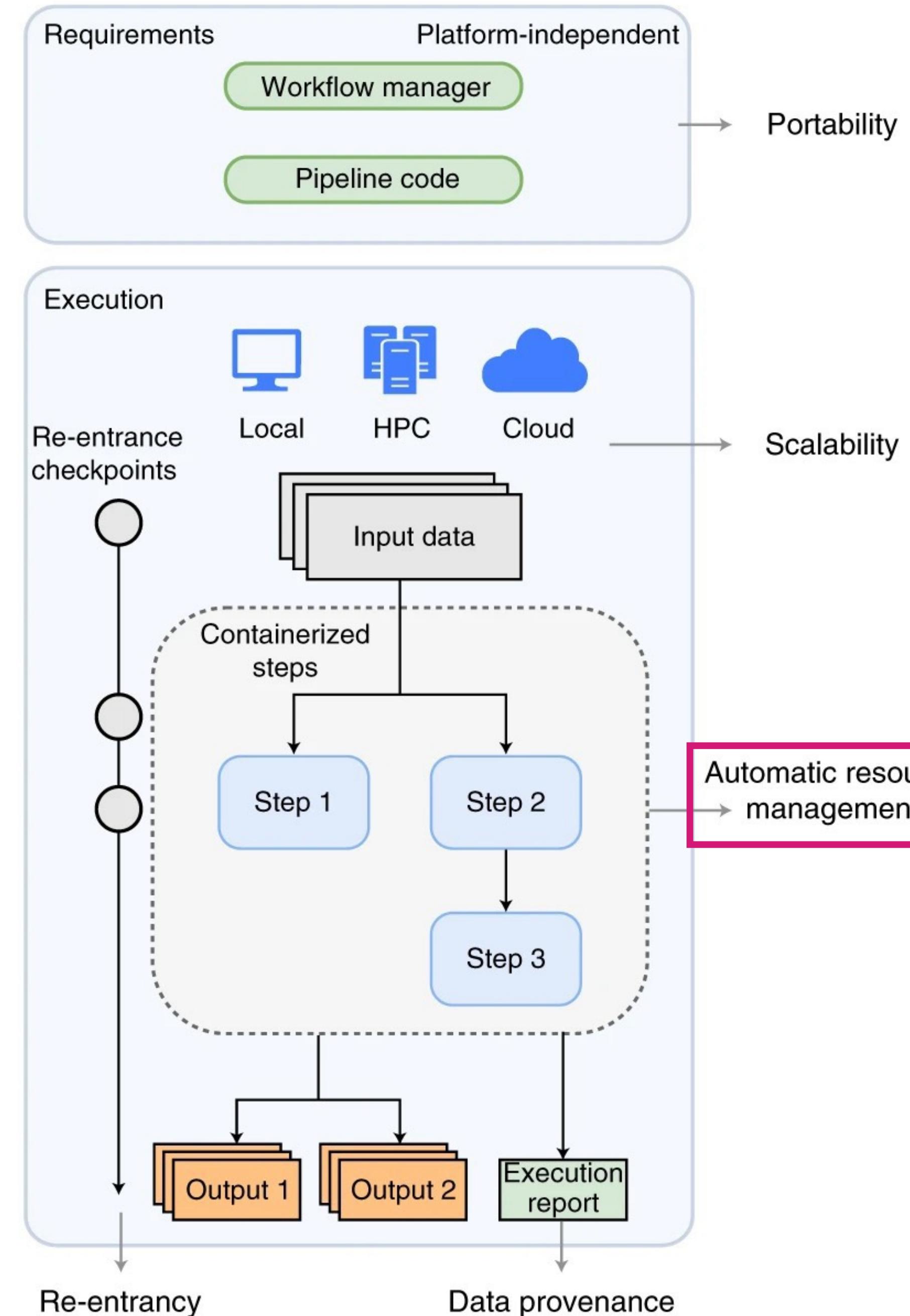
Use these techniques for increased reproducibility!



Scalability

ability to handle arbitrary amount of input data

- support for local computer, high performance computing (HPC), cloud computing
- parallelisation



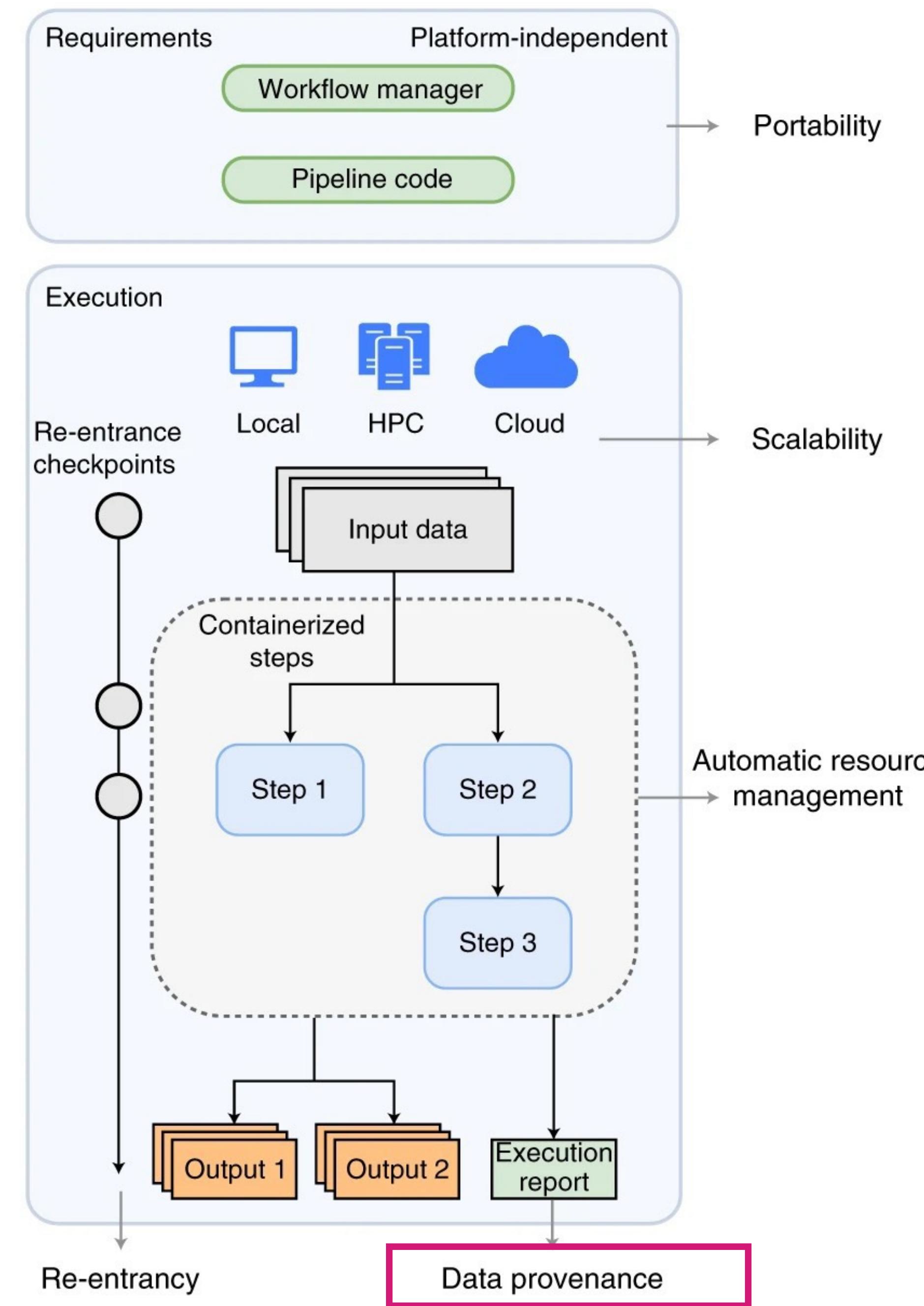
Resource Management

Allocating appropriate resources for each step

- sufficient but not excessive CPU / memory / time
- handling job rerun
- parallelisation



Effective handling of arbitrarily large datasets to minimise bottlenecks and decrease total running time.

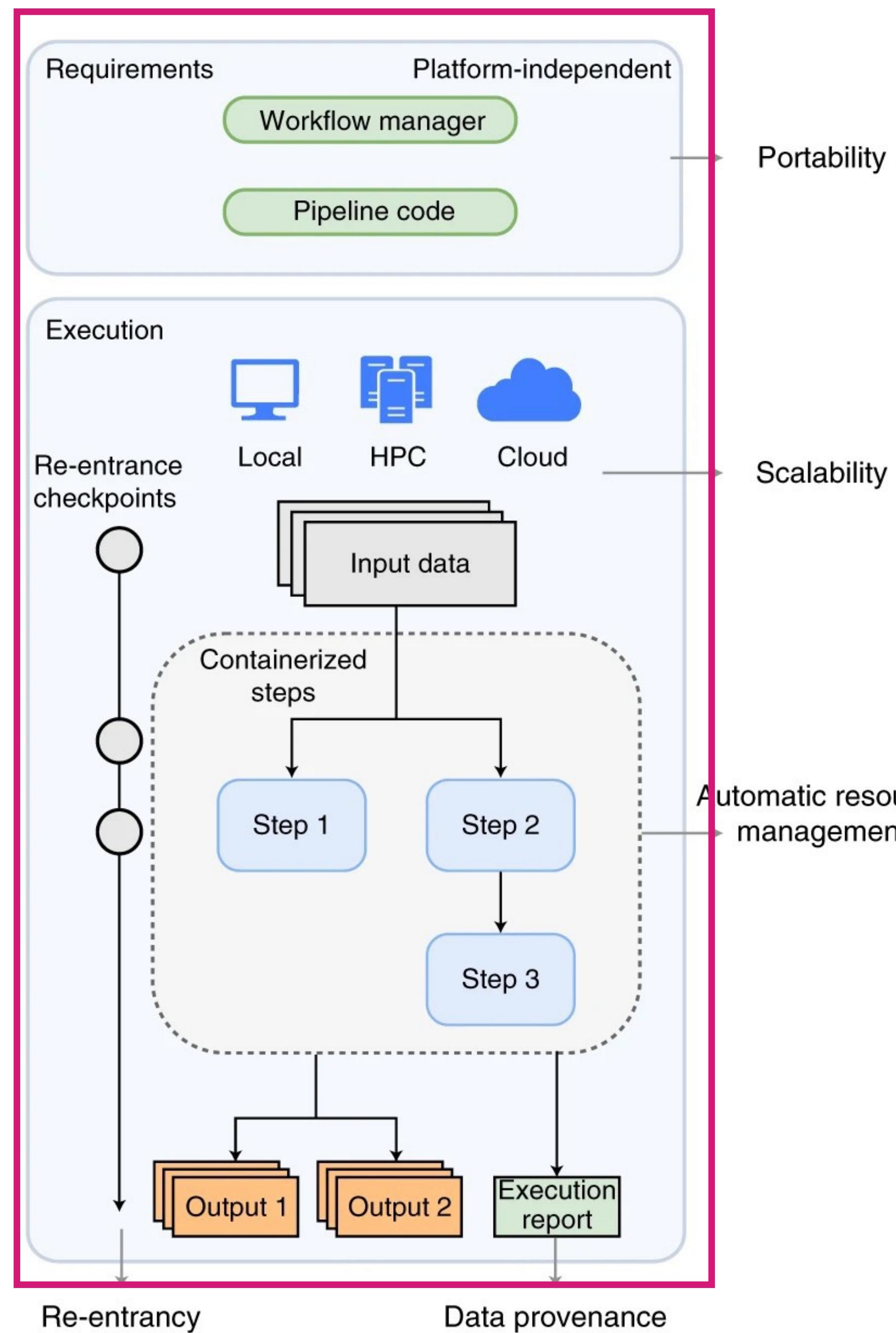


Data Provenance

Tracking of

- input data
- references
- annotations
- software versions
- parameters
- workflow performance

Documentation! Run time reports!



Workflow

The workflow itself can be versioned.



Workflow Manager Software

Domain-Specific Language (DSL)

GUI tools

nextflow



Bpipe

Luigi



Web-based open-source platform for bioinformatic workflows

- Designed for biologists, simple to use with own or public data; not suitable for sensitive patient data
- Major software available through Tool Shed ("app store", > 8k tools implemented); sometimes not all parameters are available
- Abundance of learning resources

nextflow



& other DSL tools

Domain-Specific Language (DSL) developed to meet specific need in a particular domain

WFM implemented in DSL

- Flexible, robust, portable, reproducible
- Incorporate existing tools, easily extensible
- Reusable modules
- Learning curve

nextflow

- Free, open source
- Documentation
- Community

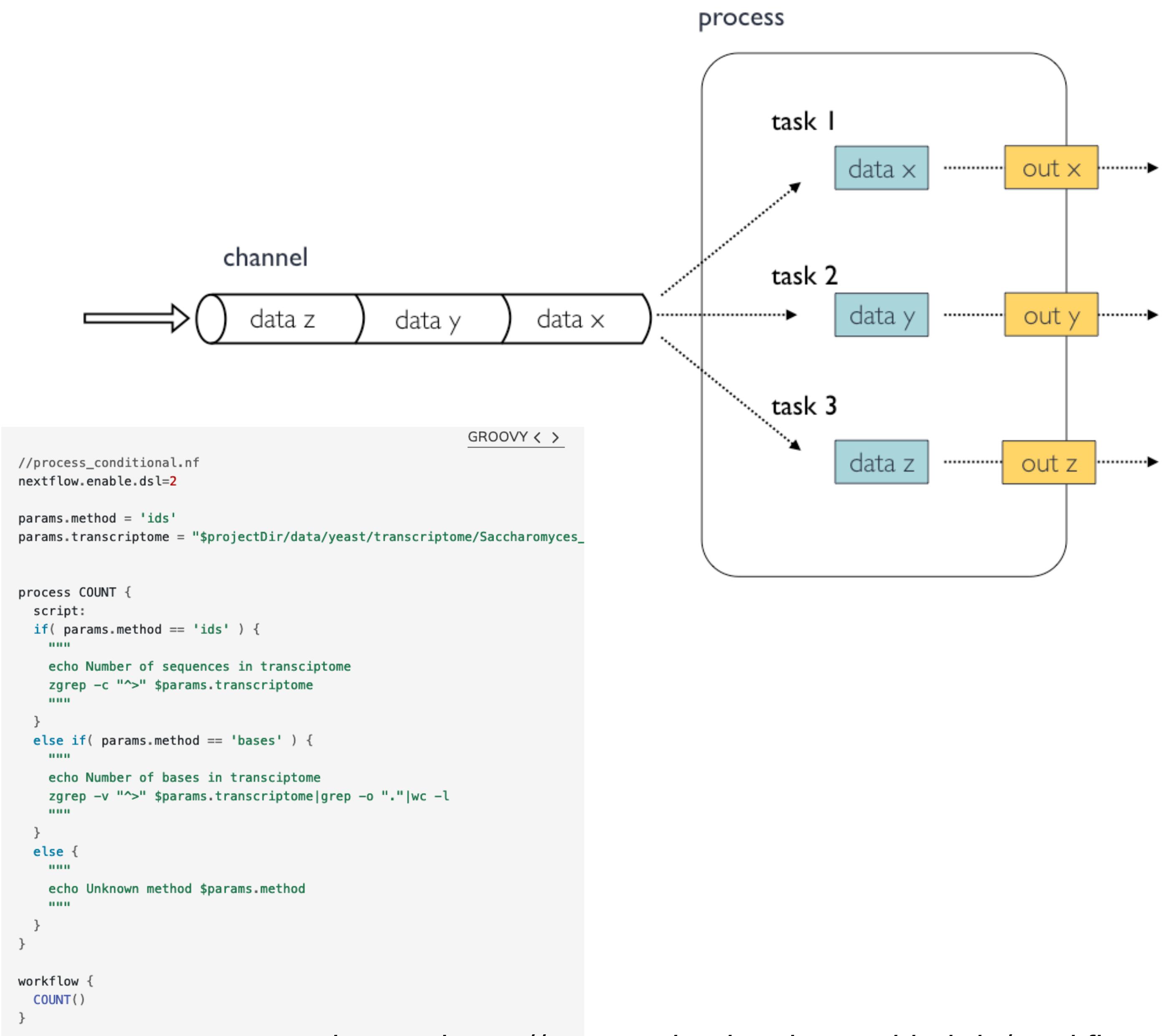
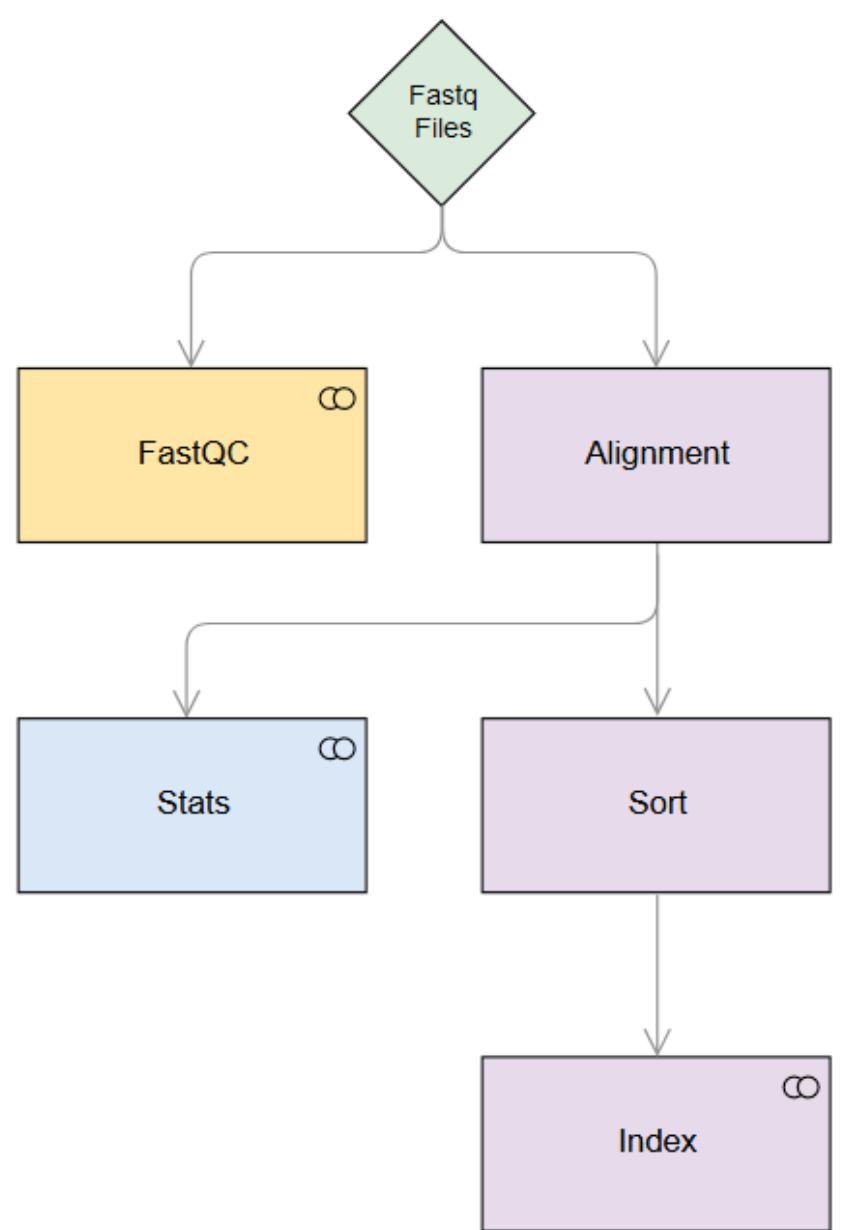
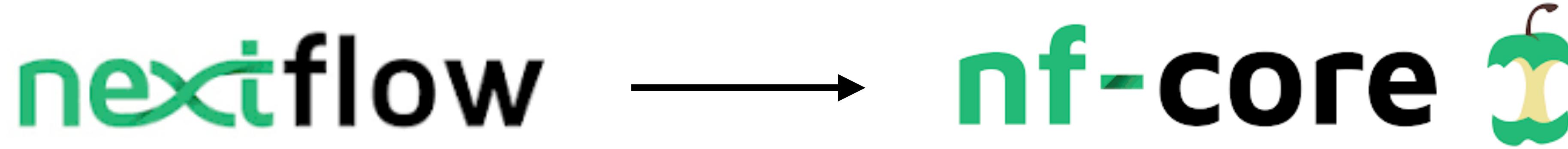


image: <https://carpentries-incubator.github.io/workflows-nextflow>



Collection of best practice pipelines for data processing

84 released pipelines

45 under development

Most genomics methods covered

methylseq, atacseq, chipseq, cutandrun, hic, mnaseseq, ...



Labs: Run nf-core pipelines on Uppmax

- to understand basic components in a Nextflow run
- to be able to find relevant nf-core pipelines



Labs: Run nf-core pipelines on Uppmax

- to understand basic components in a Nextflow run
- to be able to find relevant nf-core pipelines

Use a workflow manager if your analysis

- use > 1 tool, on > 1 data set
- contains long calculation steps
- will be shared
- requires many software dependencies

Many tools - how to make the most of them?

nextflow

CONDA


docker

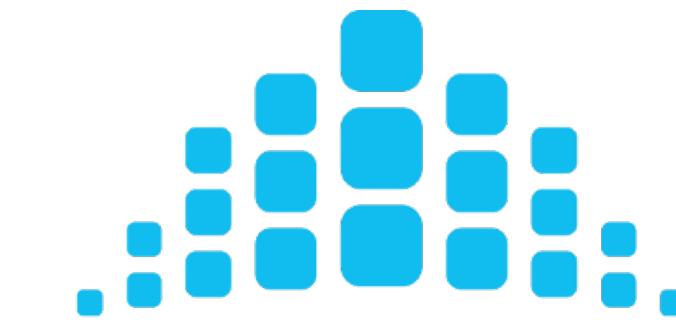

APPTAINER


GitHub


GitLab






slurm
workload manager


Bitbucket

Many tools - how to make the most of them?



[Talk to us](#) [Services](#) [Training](#) [About us](#) [Contact](#)



Tools for Reproducible Research

Description

One of the key principles of proper scientific procedure is the act of repeating an experiment or analysis and being able to reach similar conclusions. Published research based on computational analysis, e.g. bioinformatics or computational biology, have often suffered from incomplete method descriptions (e.g. list of used software versions); unavailable raw data; and incomplete, undocumented and/or unavailable code. This essentially prevents any possibility of attempting to reproduce the results of such studies. The term “reproducible research” has been used to describe the idea that a scientific publication based on computational analysis should be distributed along with all the raw data and metadata used in the study, all the code and/



nextflow

GitHub



APP TAINER

Thank you for listening

snakemake

slurm
workload manager

CONDA

GitLab

Bitbucket

docker