Workflow Management

Epigenomics Data Analysis Workshop

Agata Smialowska 2024

Scientific Data Analysis

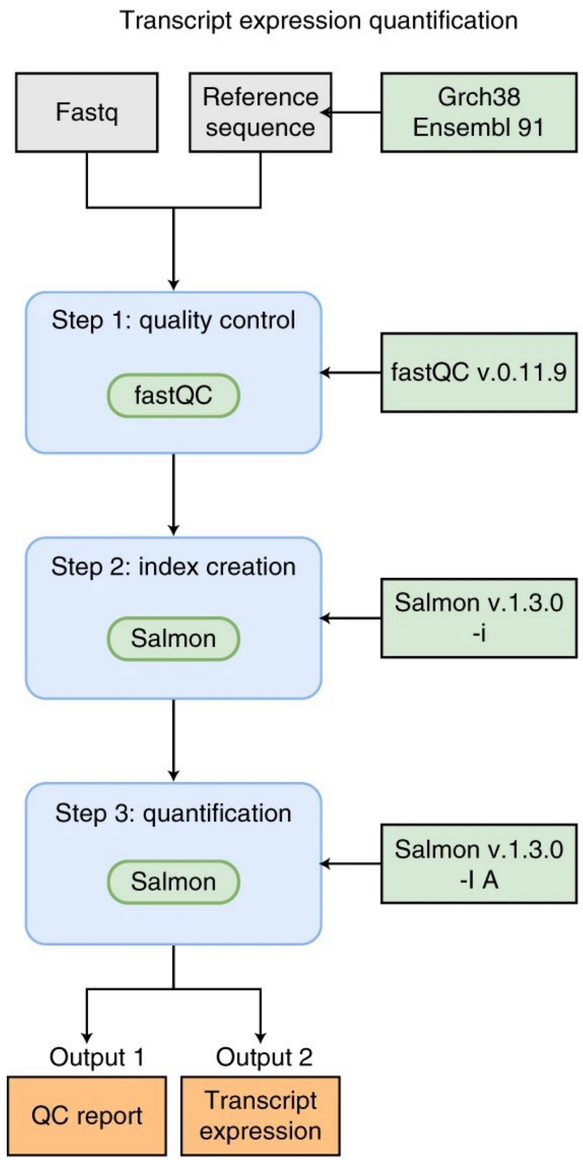
Perspective | Published: 23 September 2021

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

Laura Wratten, Andreas Wilm & Jonathan Göke

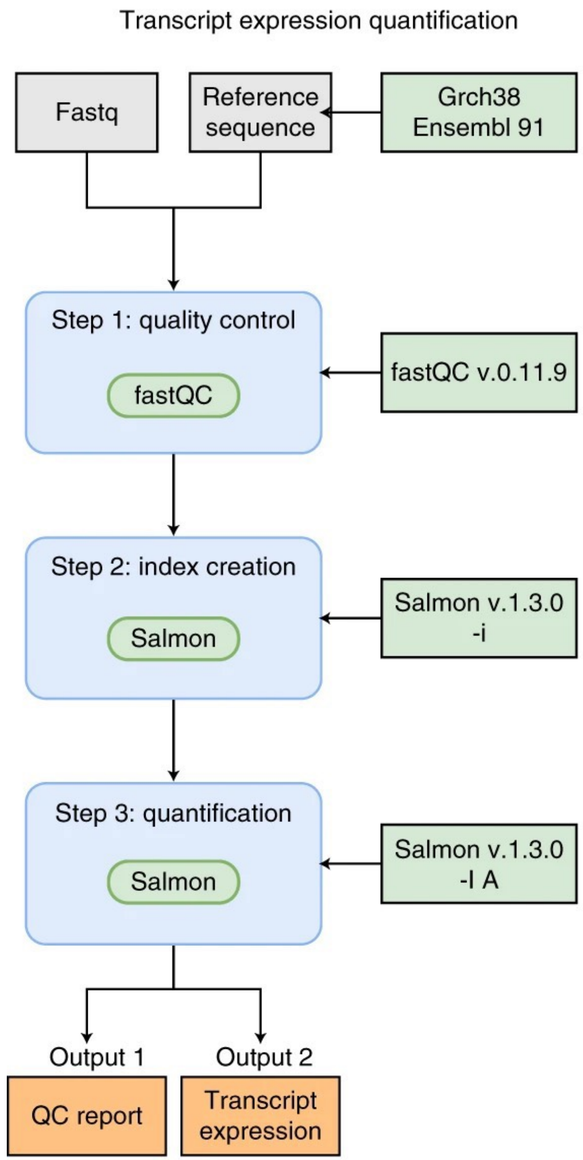
Nature Methods **18**, 1161–1168 (2021) | Cite this article

Scientific Data Analysis



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

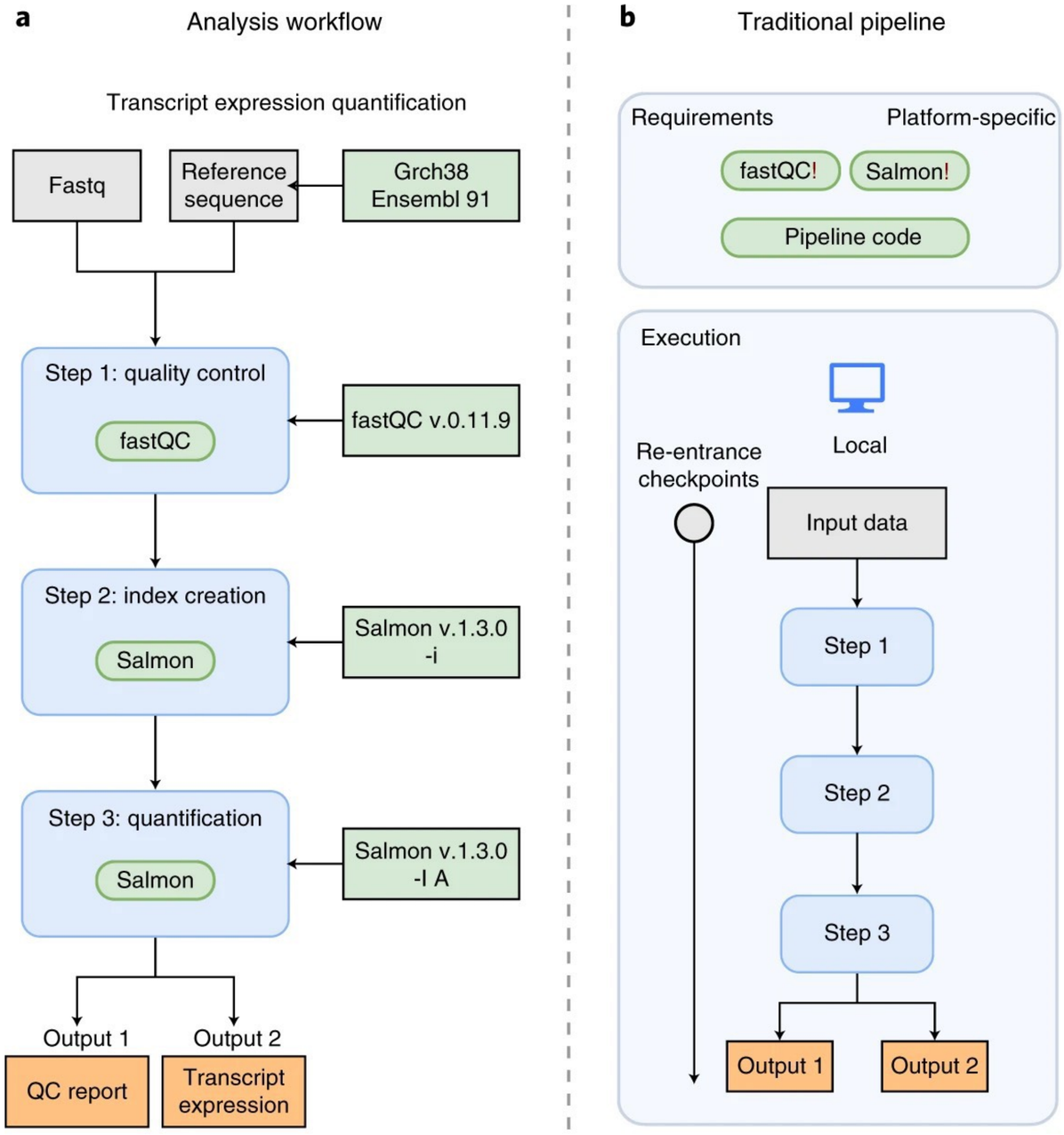
Scientific Data Analysis



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

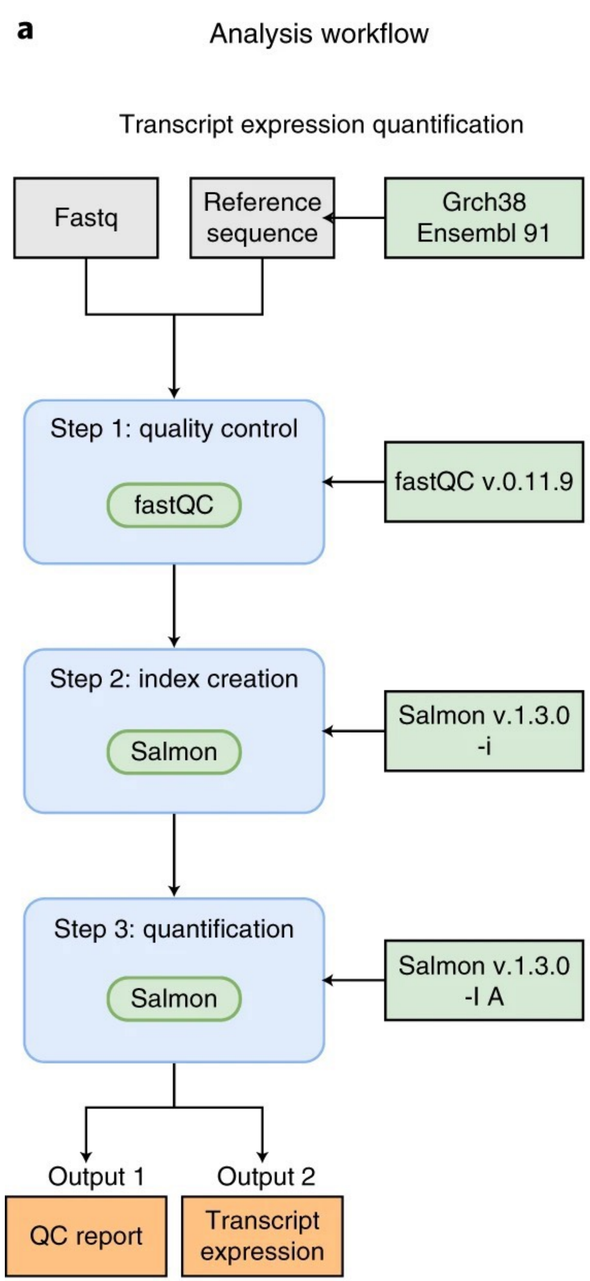
Not trivial

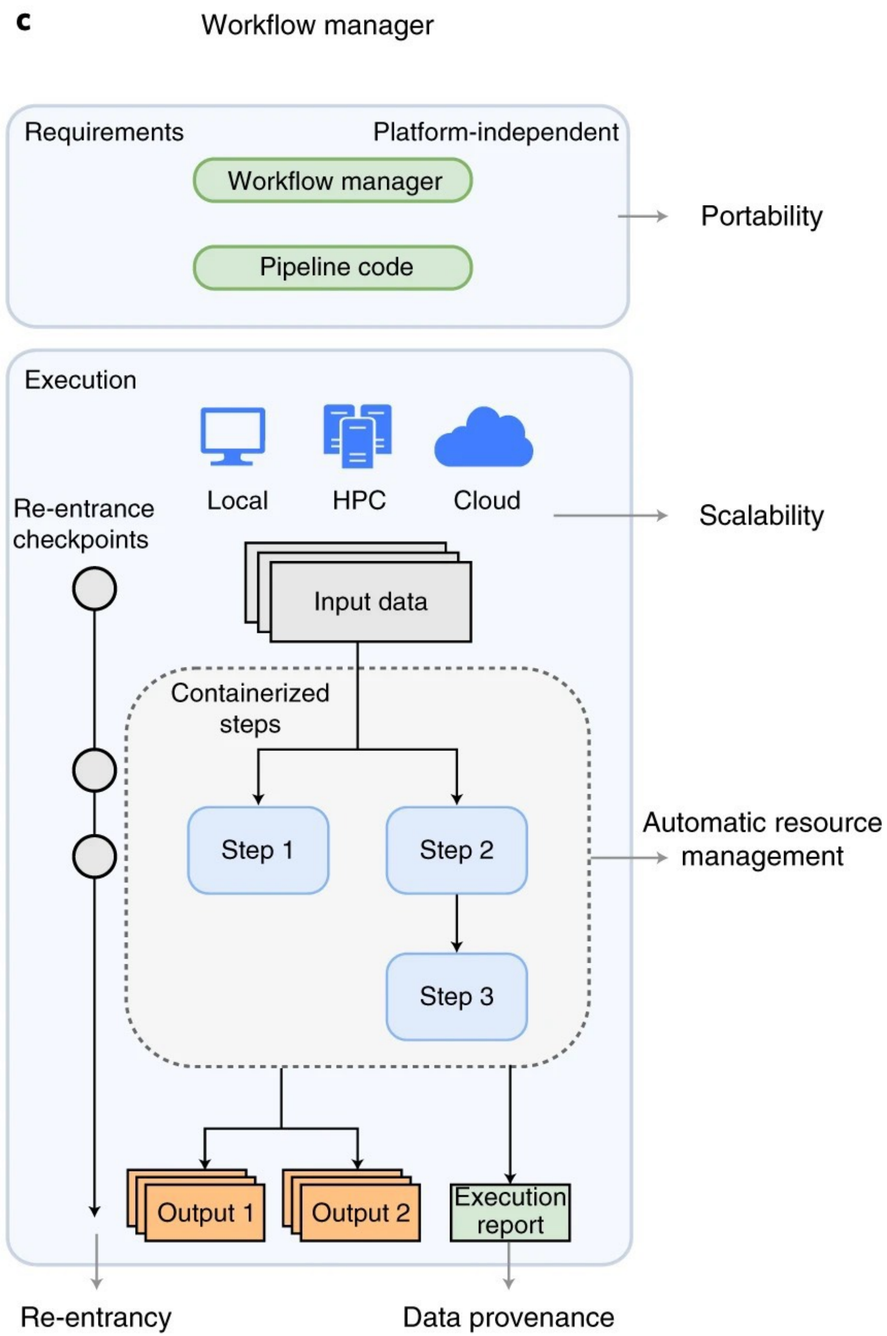
- Analysis reruns
- Large workflows
- Collaboration
- <u>Reproducibility</u>



Bash scripts to chain multiple tools to make the analyses easier

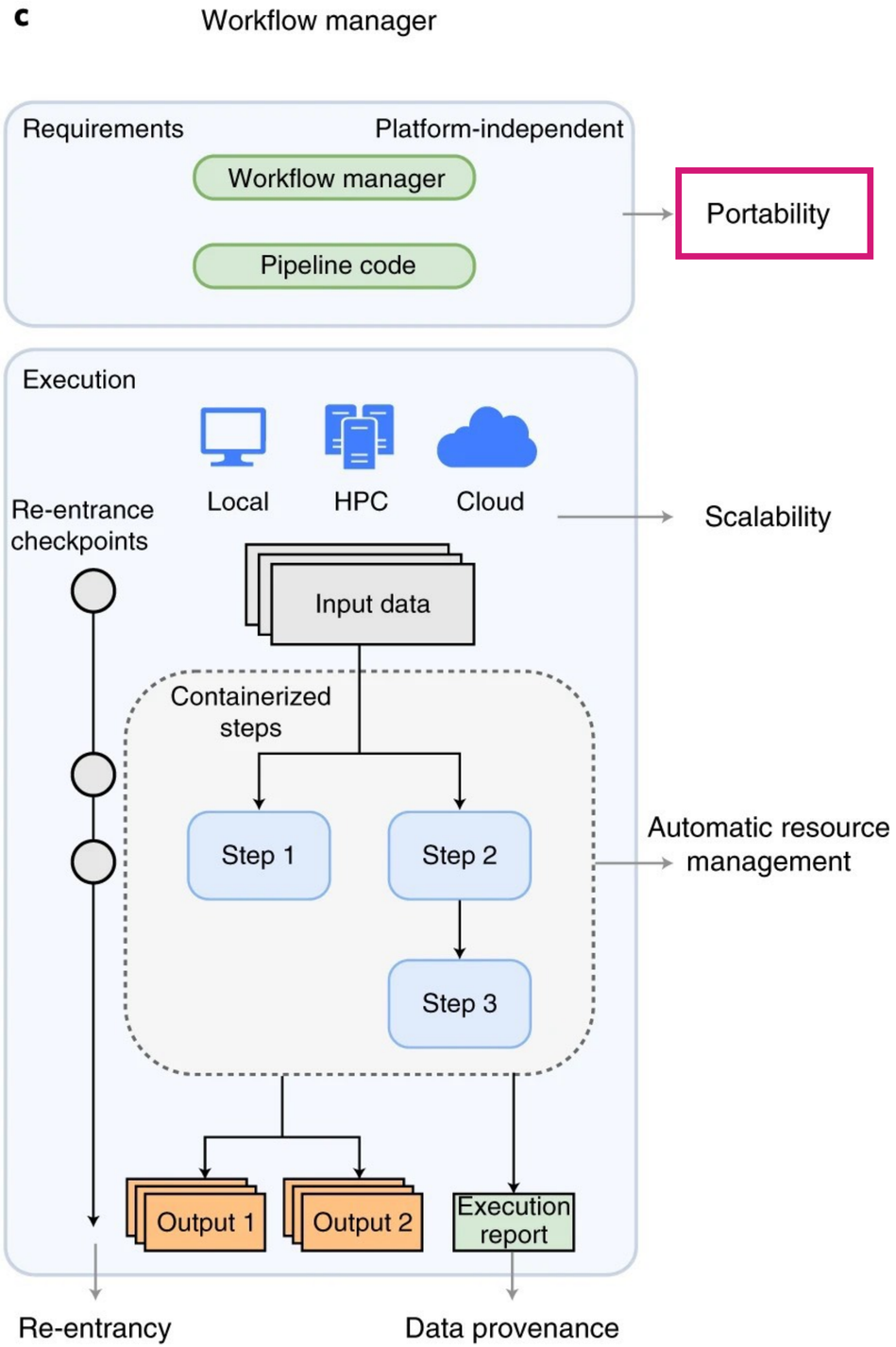
- Cannot resume failed run
- Connected to local infrastructure
- Lack of documentation
  - parameters
  - versions
- Low portability
- Cumbersome maintenance
  - tools
  - workflow





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

execution across different platforms and over time

- Package Managers
  - automate installing and manage software dependencies

CONDA

- Software containers
  - isolated execution environment

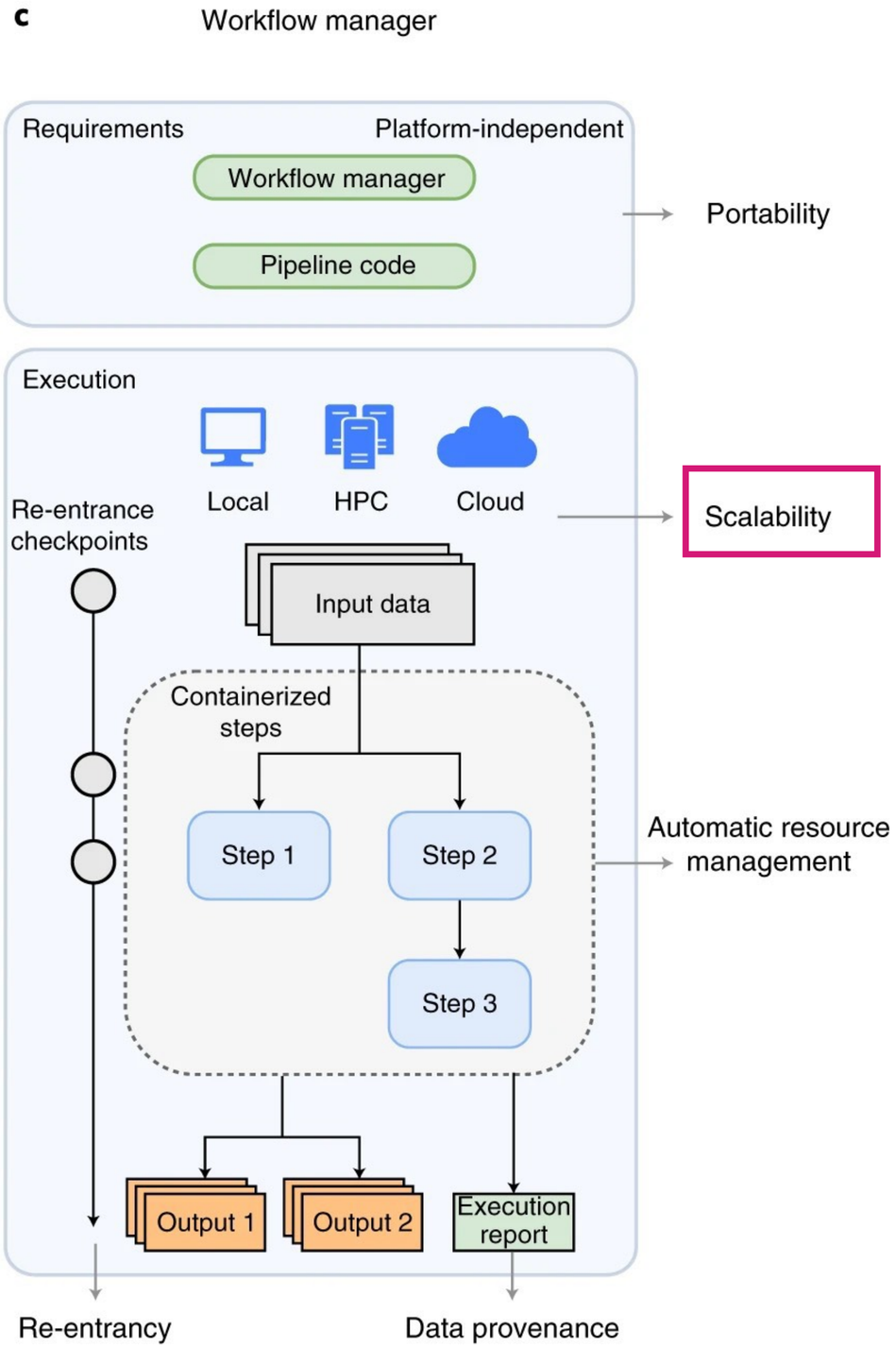
APPTAINER

docker

  - BioContainers: prebuilt containers of all major tools

OS and software versions can affect the results.

Use these techniques for increased reproducibility!



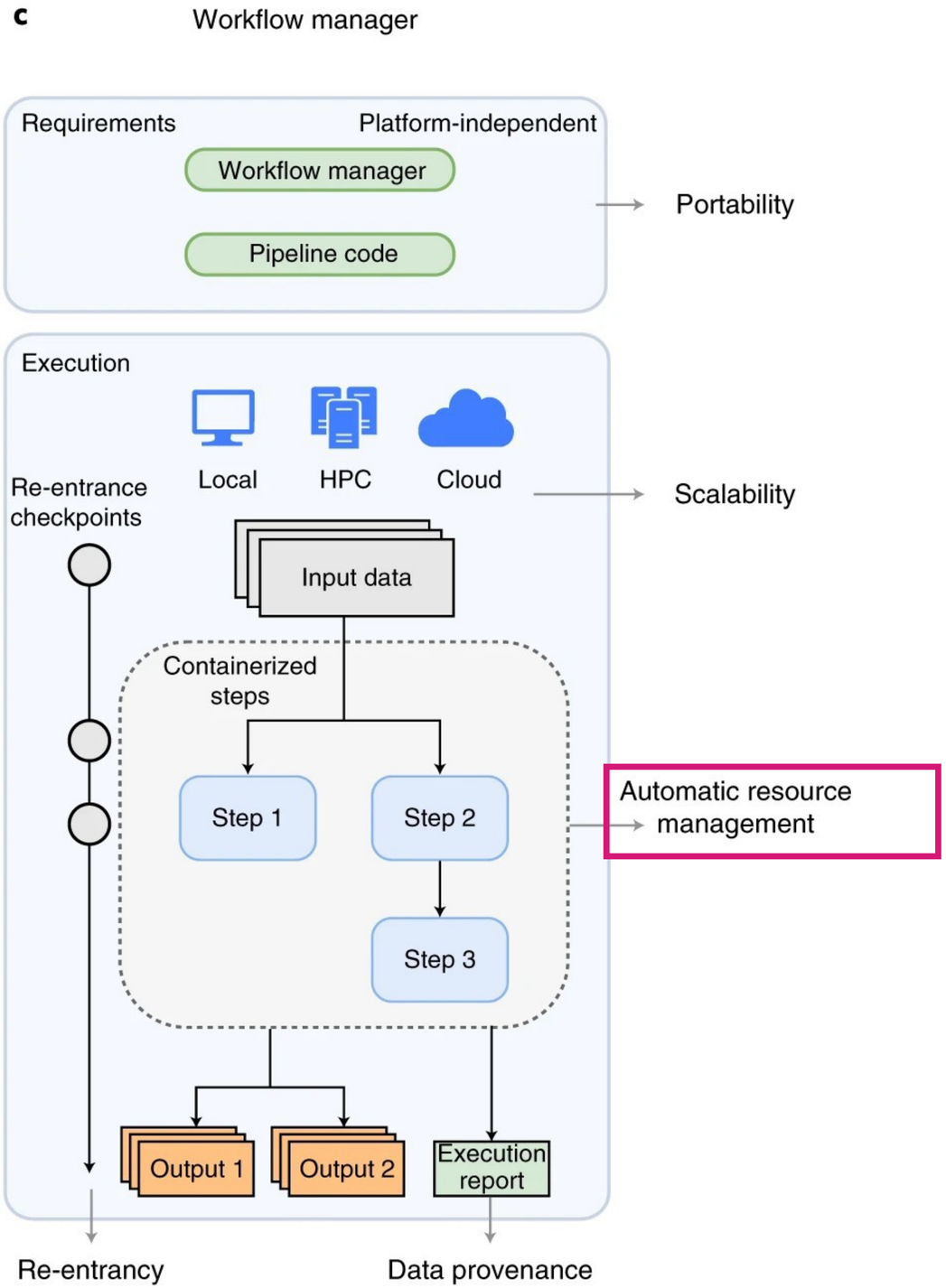
Scalability

ability to handle arbitrary amount of input data

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

docker

APPTAINER



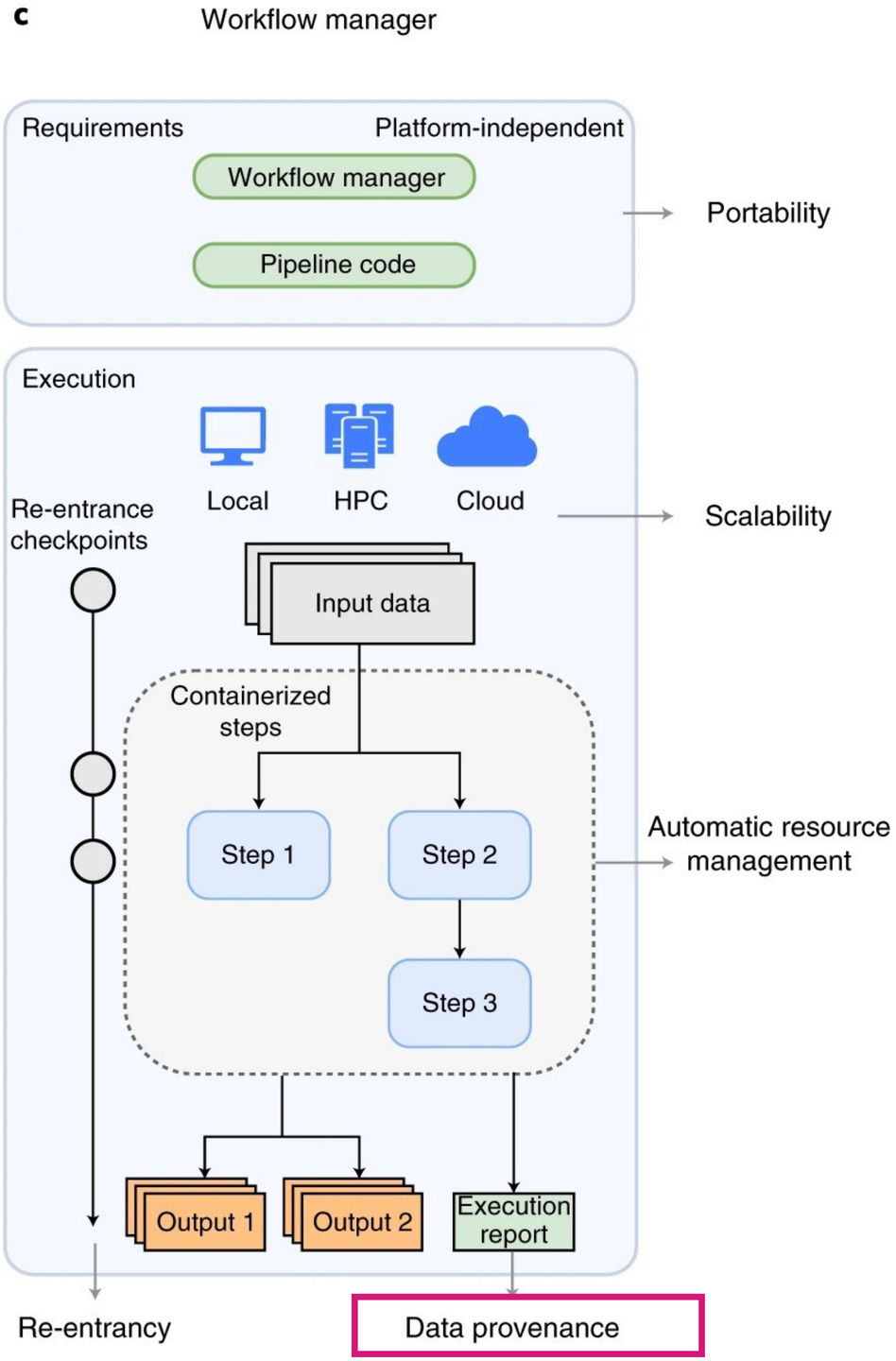
Resource Management

Allocating appropriate resources for each step

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

slurm
workload manager

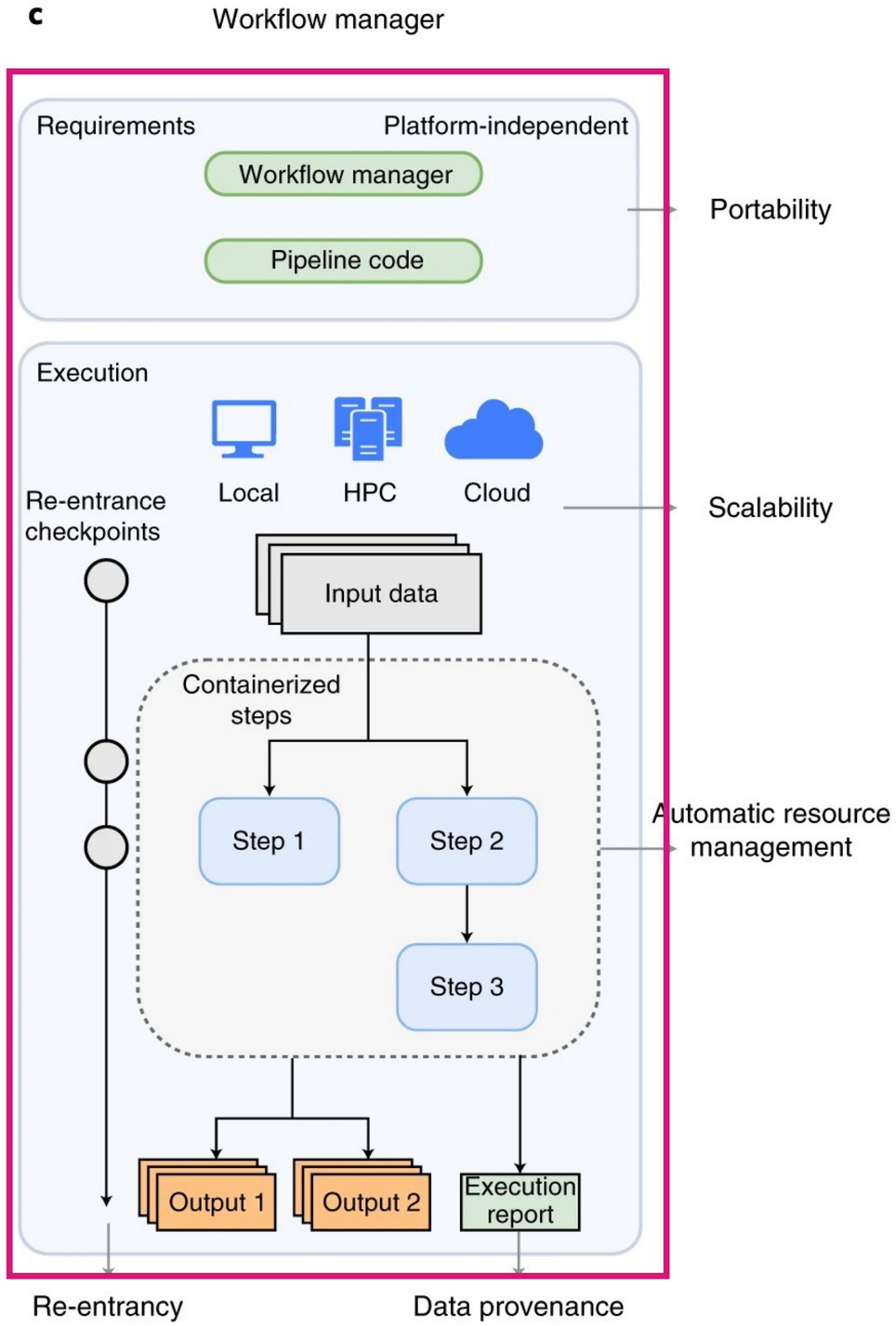
Effective handling of arbitrarily large datasets to minimise bottlenecks and decrease total running time.



Data Provenance

Any change in input data, references, annotations, software versions, parameters is tracked and recorded.

Documentation! Run time reports!



Workflow

The workflow itself can be versioned.

GitHub

GitLab

Bitbucket

Workflow Manager Software

Domain-Specific Language (DSL)

nextflow

snakemake

COMMON
WORKFLOW
LANGUAGE

Bpipe

Luigi

GUI tools

Galaxy
PROJECT

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 snakemake & other DSL tools

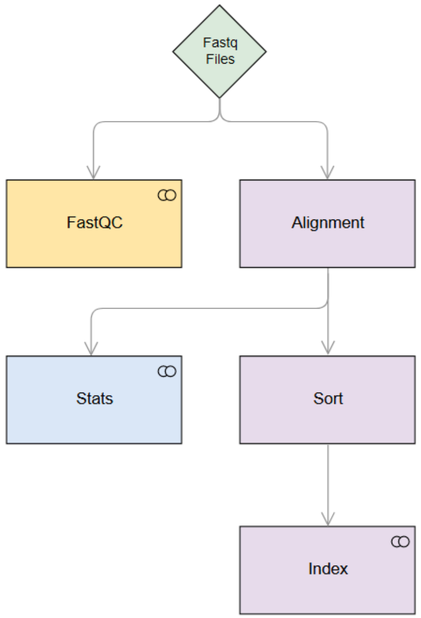
Domain-Specific Language (DSL) developed to meet specific need in a particular domain

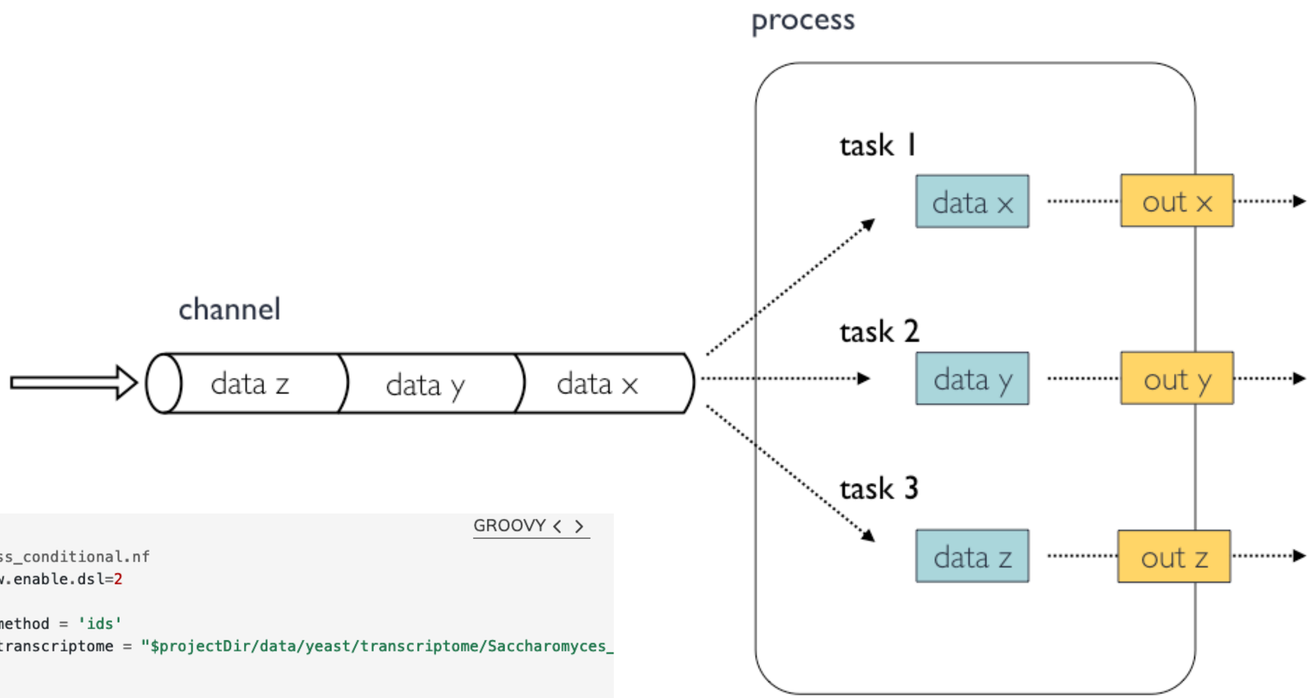
WFM implemented in DSL

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

nextflow

- Free, open source
- Documentation
- Community





GROOVY

```
//process_conditional.nf
nextflow.enable.dsl=2

params.method = 'ids'
params.transcriptome = "$projectDir/data/yeast/transcriptome/Saccharomyces_


process COUNT {
  script:
  if( params.method == 'ids' ) {
    """
    echo Number of sequences in transciptome
    zgrep -c "^>" $params.transcriptome
    """
  }
  else if( params.method == 'bases' ) {
    """
    echo Number of bases in transciptome
    zgrep -v "^>" $params.transcriptome|grep -o "."|wc -l
    """
  }
  else {
    """
    echo Unknown method $params.method
    """
  }
}

workflow {
  COUNT()
}
```

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

nextflow → nf-core

Collection of best practice pipelines for data processing

66 released pipelines

33 under development

Most genomics methods covered

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

nf-core

Labs: Run nf-core pipelines on Uppmax

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

nf-core

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

snakemake

slurm
workload manager

Bitbucket