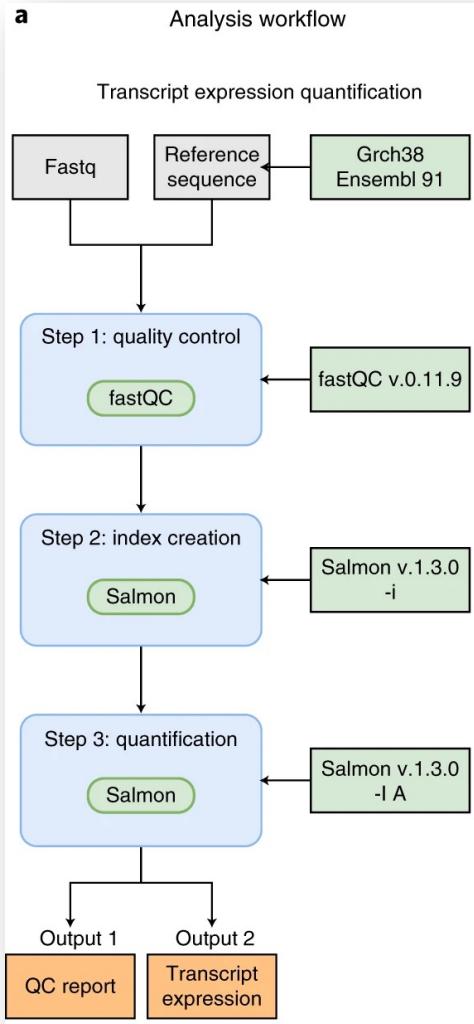


Workflow Management

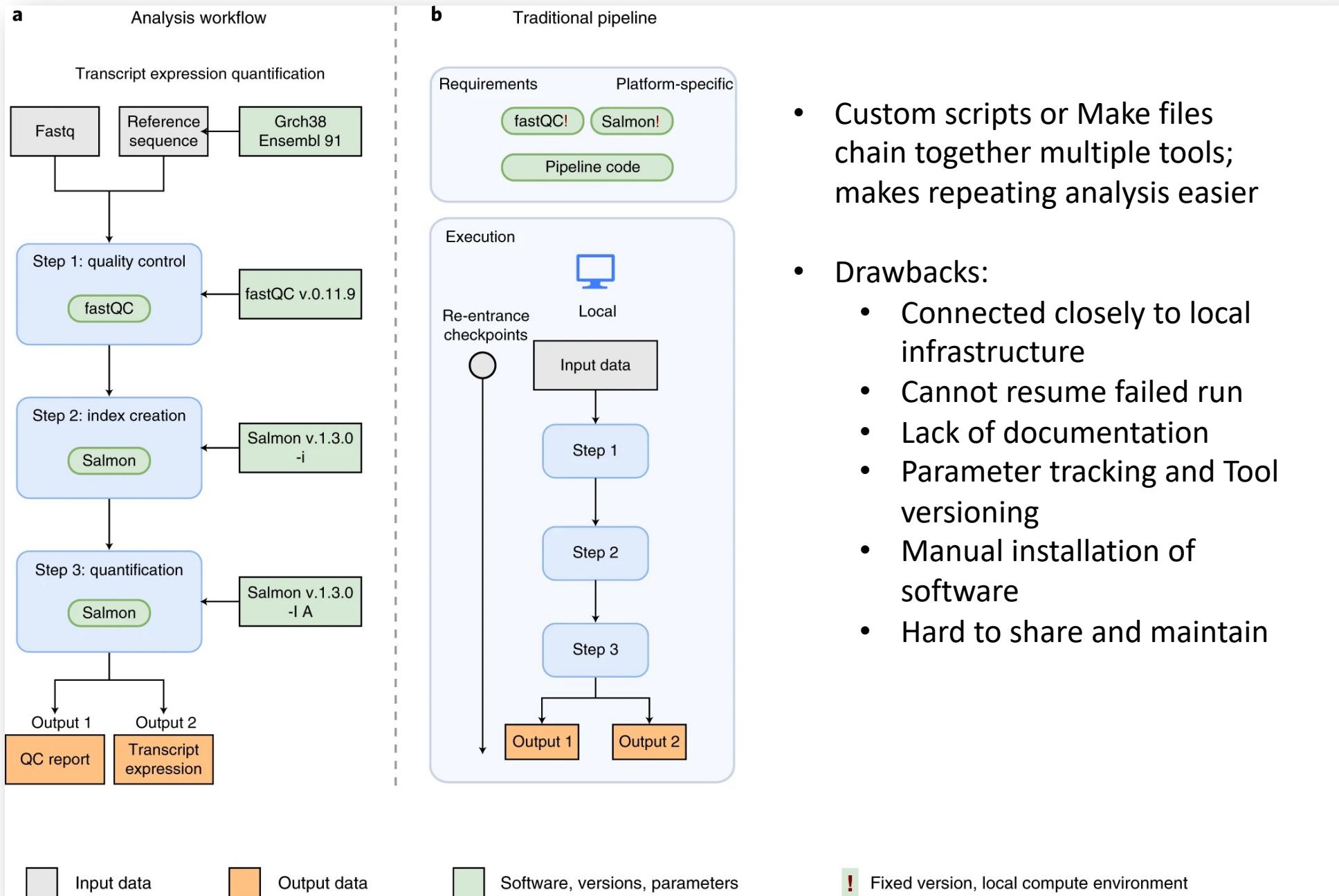
Epigenomics Data Analysis 2021

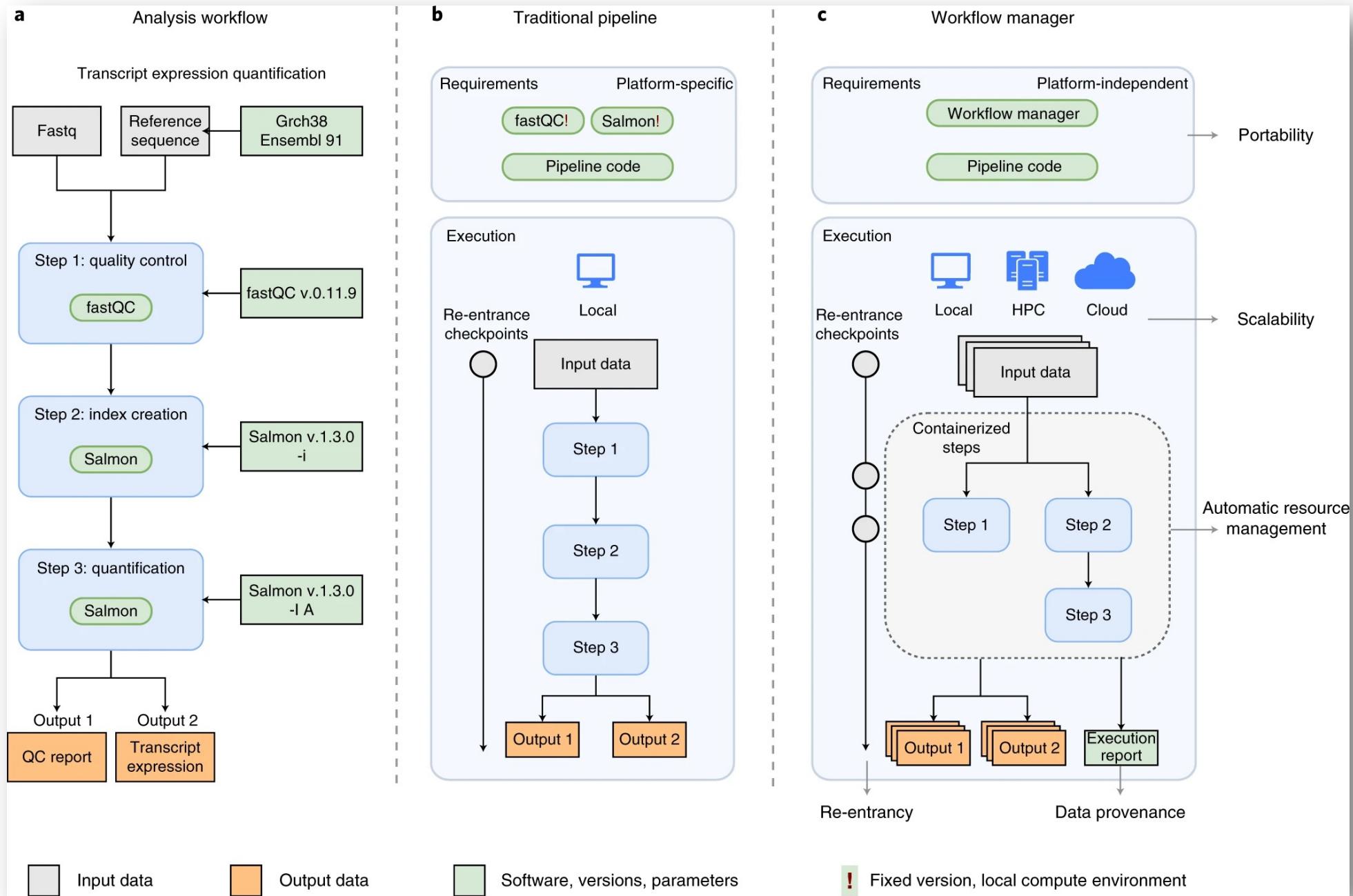
Scientific Data Analysis in practice

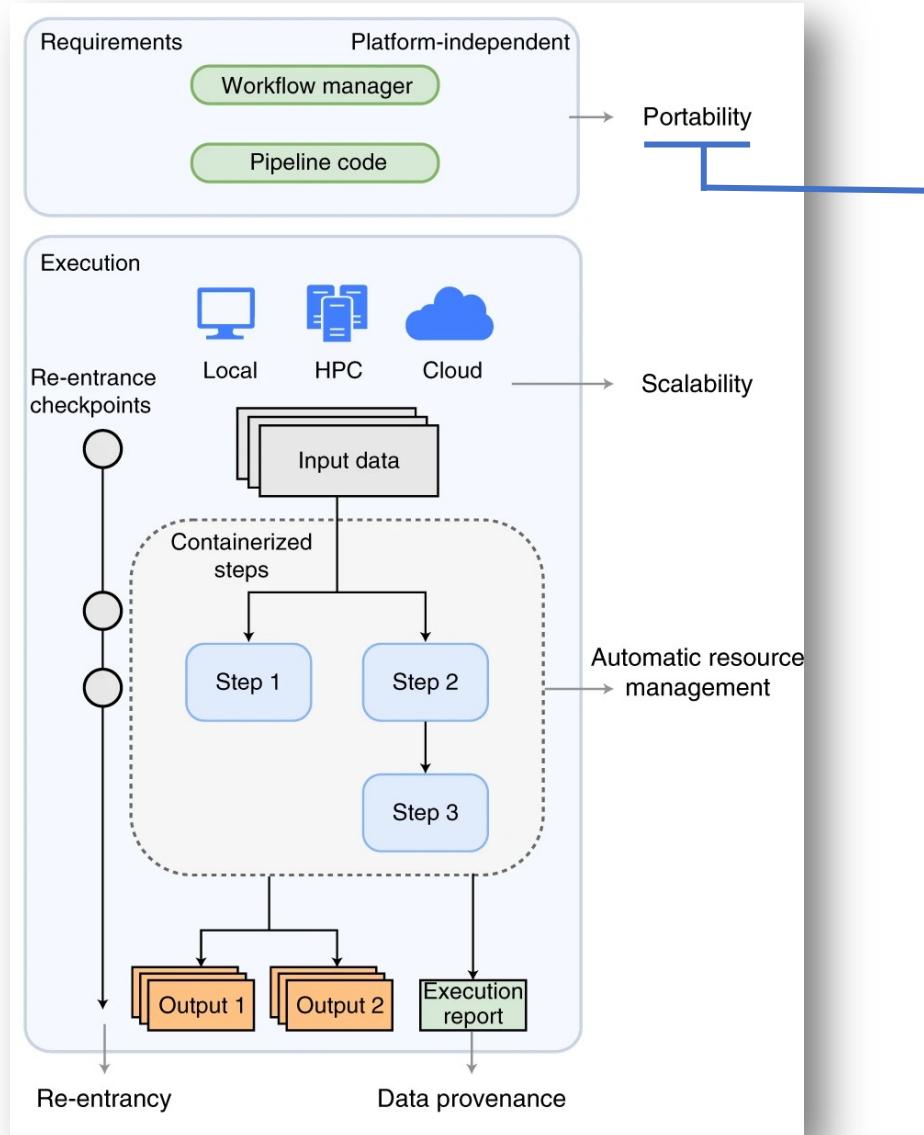


- Often no single tool; many tools need to be combined
- References, databases and software versioned
- Every tool comes with their own set of parameters
- Experimentation is crucial; reruns necessary
- Reproducibility becomes an issue
- Rerun part of the analysis
- Resource management?
- ...



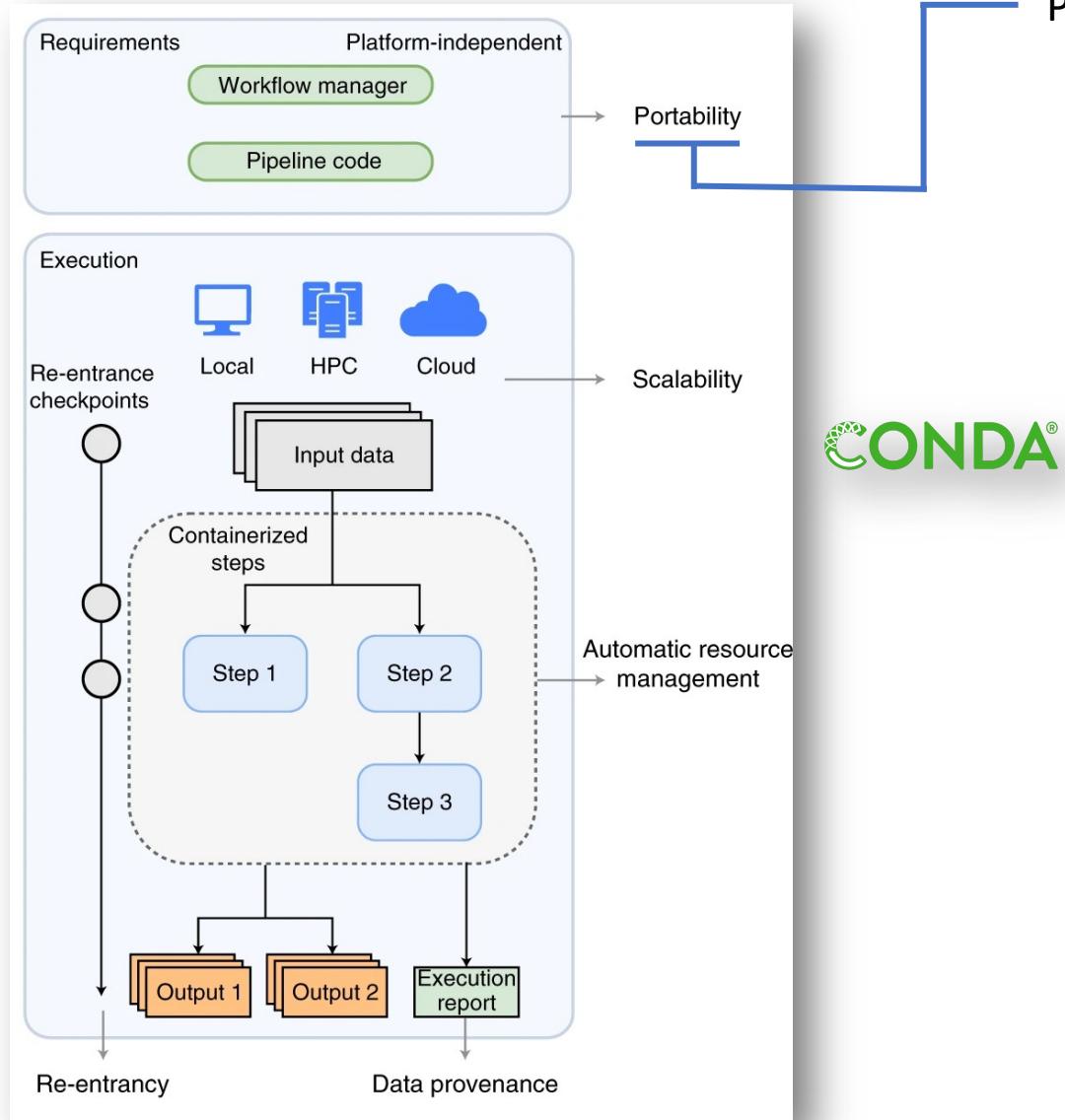






Portability

Execution with the same functionality across different platforms and over time.

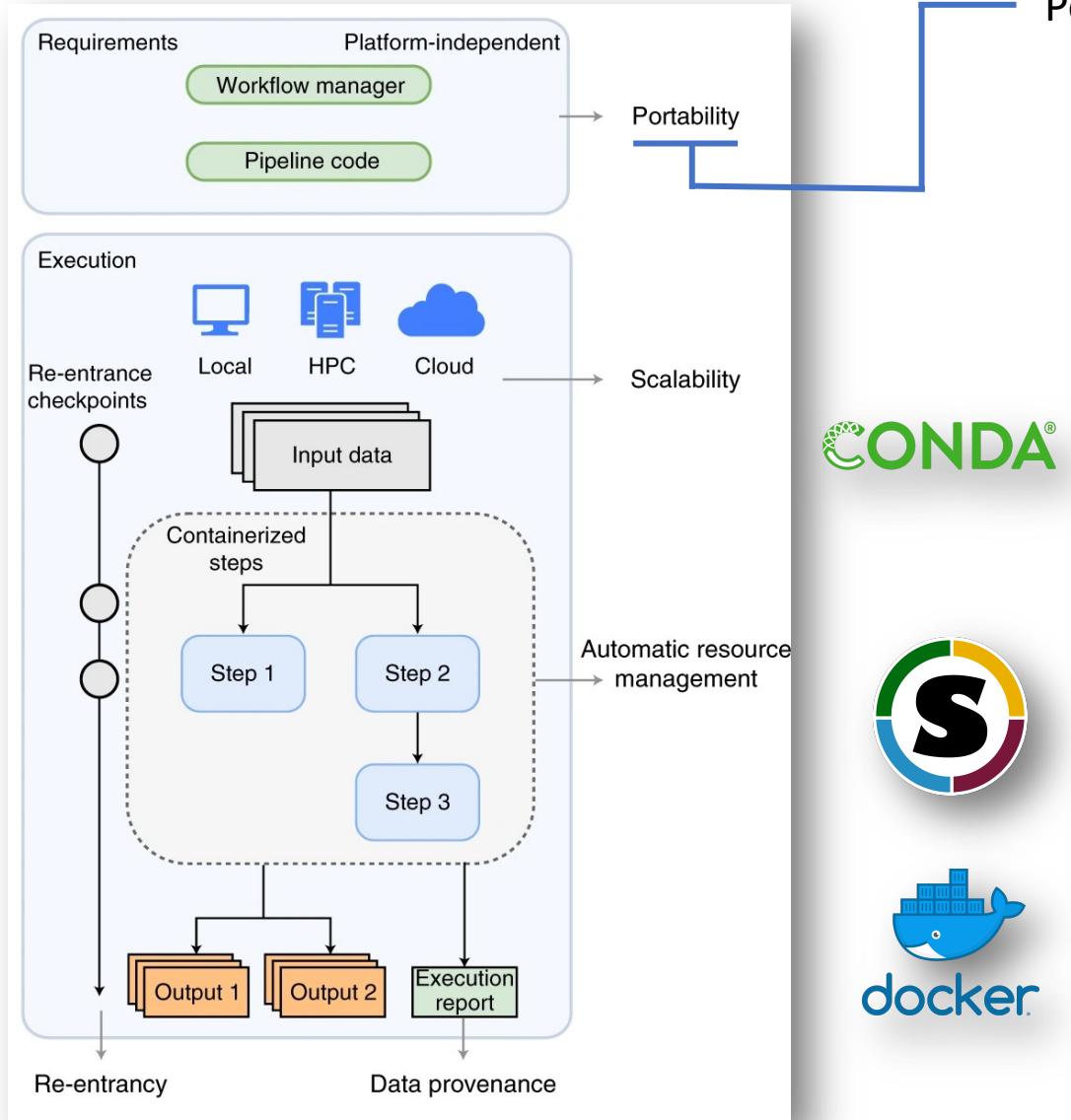


Portability

Execution with the same functionality across different platforms and over time:

- **Package Managers**
 - Automate the process of installing and configuring software
 - Obtain all tools and dependencies with a single command
 - Homebrew, conda, ...

CONDA®



Portability

Execution with the same functionality across different platforms and over time:

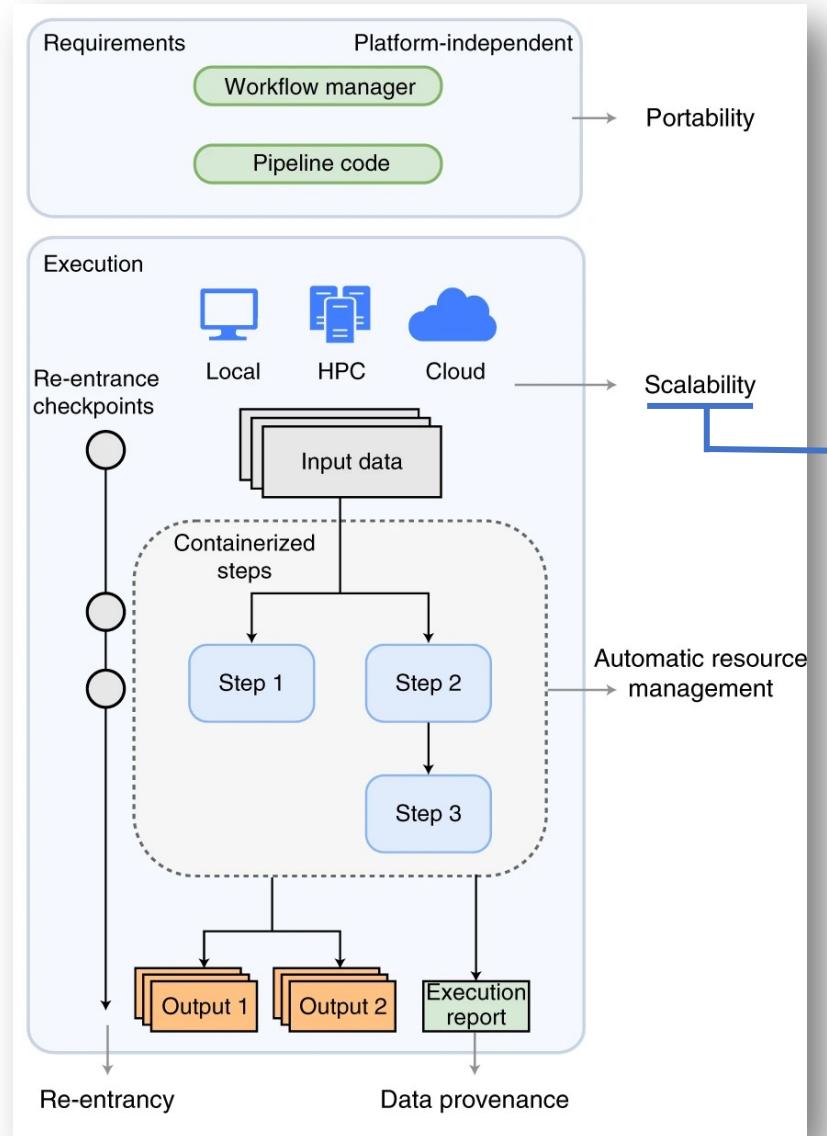
- **Package Managers**
 - Automate the process of installing and configuring software
 - Obtain all tools and dependencies with a single command
 - Homebrew, conda, ...
- **Containerization Software**
 - “Lightweight” virtualization technology
 - Combines software, dependencies and operating system in a self-contained “package”
 - Docker, Singularity
 - BioContainers: prebuilt containers 1000s of tools

CONDA



docker

OS & software version can influence results; use these techniques for increased reproducibility!

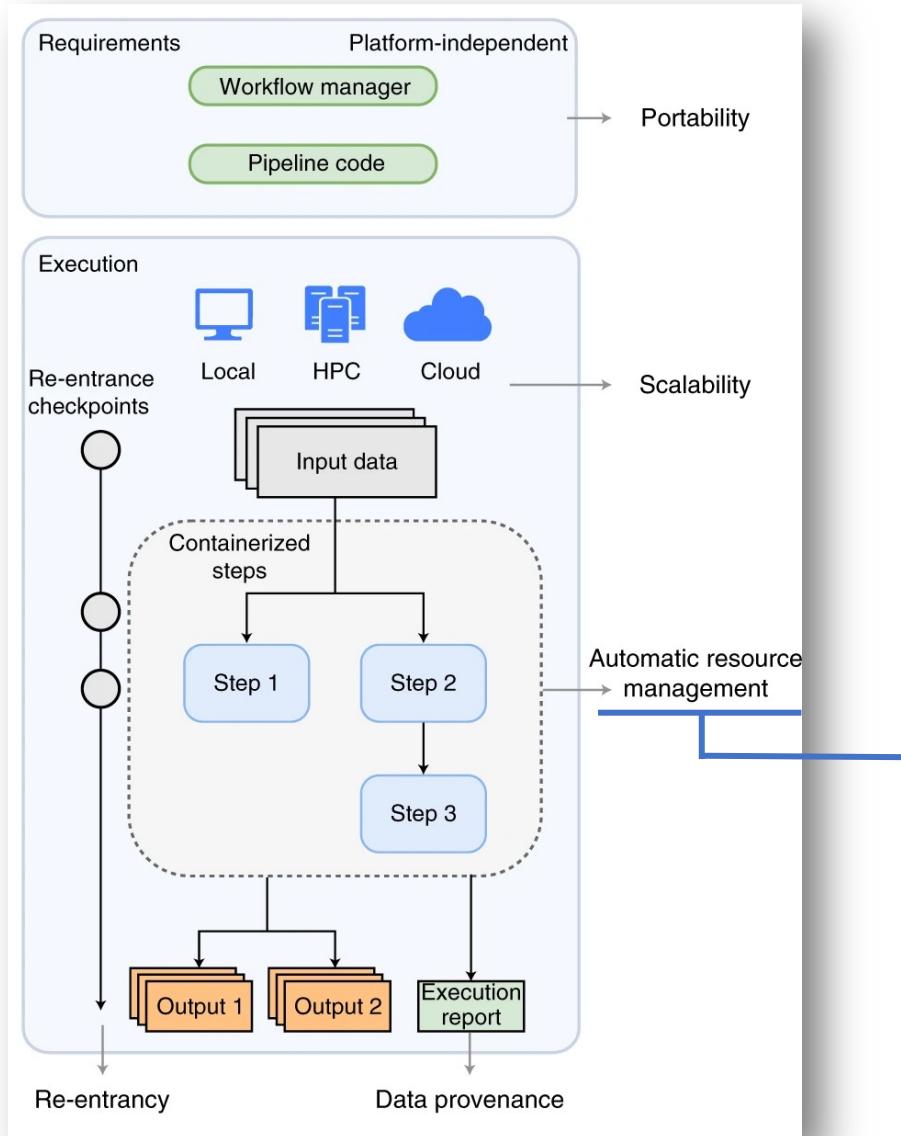


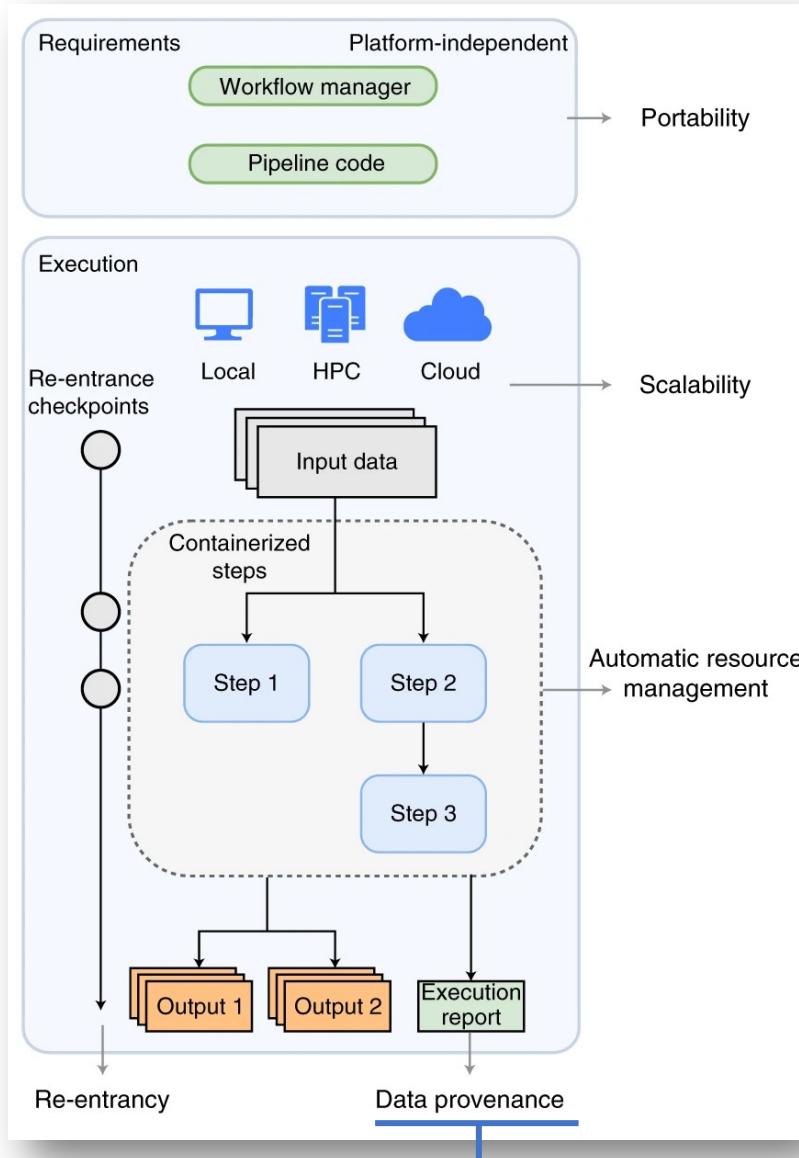
Scalability

Ability to handle any size and quantity of input data:

- Built-in support for local computer, high-performance computing and cloud computing.

Choose the right execution infrastructure to enable efficient analysis of small and large datasets!



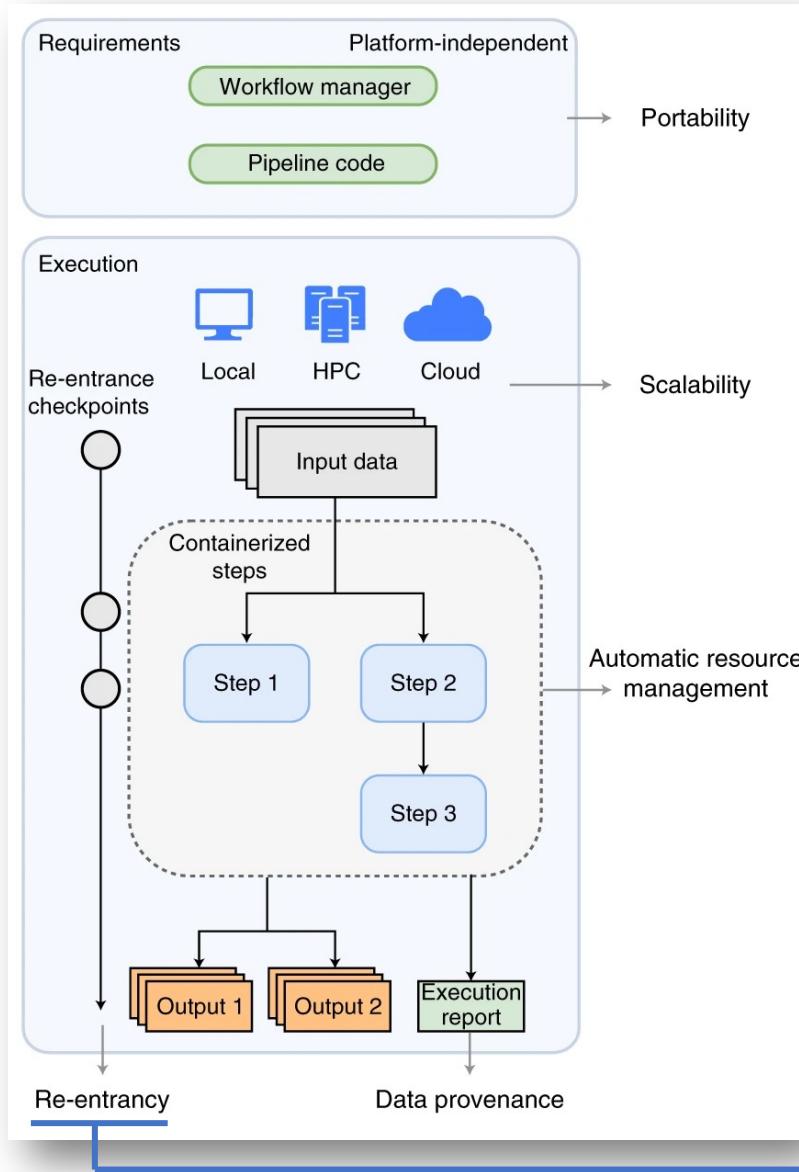


Data provenance

Any change in software version, parameters or reference annotation version can alter results:

- WFM automates tracking of input parameters and tool versions
- Detailed information in execution report
- Workflow itself can be archived and versioned

High level documentation enables transparency, code sharing and long-term reproducibility!



Re-entrancy

Run a pipeline from its last successfully executed step, rather than from the beginning:

- Error in pipeline; only repeat necessary steps
- Avoid recalculation of very long steps
- Cache intermediate results and data files
- Save significant time and resources

Re-entrancy saves significant time and compute resources and is a **key advantage** of workflow managers!

Rapid rise in manager software...

The three technologies bioinformaticians need to be using right now

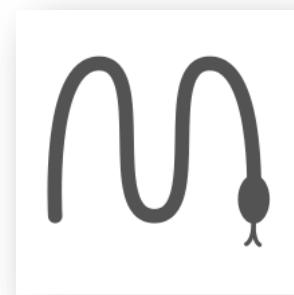
12TH AUGUST 2019 / BIOMICKWATSON / COMMENTS OFF

2. Workflow management systems

It's a law of the universe that every bioinformatician will say they have a "pipeline", but often I find this is actually a bash, perl or python script. They were great for their time, but it is time to move on.

Rapid rise in manager software...

- Many workflow management systems have been developed



nextflow

= Galaxy
PROJECT

Luigi

... and hundreds more

Three types of Workflow Managers

- Graphical Interface
 - No programming experience needed; drag and drop
- Domain-Specific Language (DSL)
 - Dedicated programming language
- Programming-library-based
 - Developed within and for a specific language



- Web-based platform for bioinformatic workflows
 - Open source; more than 7,500 citations
- Designed for biologists to work with their own and public data
- “App store” for bioinformatic tools
 - 8140 tools (<https://toolshed.g2.bx.psu.edu>)
- Abundance of learning resources



File and metatools

Get data: USCS, Uniprot
Send data: GenomeSpace Exporter
Convert

Text tools

Text manipulation
Filter and Sort
Join, Subtract and Group
Statistics

Annotation, ontologies

SortMeRNA, Aragorn, Roary, Prokka, Augustus, KOBAS, Glimmer, antiSMASH, etc

Genomics, HTS

Quality control: FastQC, MultiQC, Trim Galore!, etc
Alignment: BLAST, Diamond, etc
Mapping: Bowtie2, STAR, HISAT2, BWA, segemehl, etc
Assembly: Unicycler, SPades, Quast, etc
Transcriptomics: FeatureCounts, DESeq2, Trinity, Salmon, etc
RNA: LocARNA, RNAfold, RNAz, RNAPlot, etc
Variant Calling: FreeBayes, Gemini, VCFTools, SnpEff, etc
Peak Calling: MACS2, Piranha, PEAKachu, etc
Epigenetics: Bismark, metilene, bwameth, MethylDackel, etc
deepTools, SAM Tools, HicExplorer, Picard, EMBOSS, etc

Metagenomics

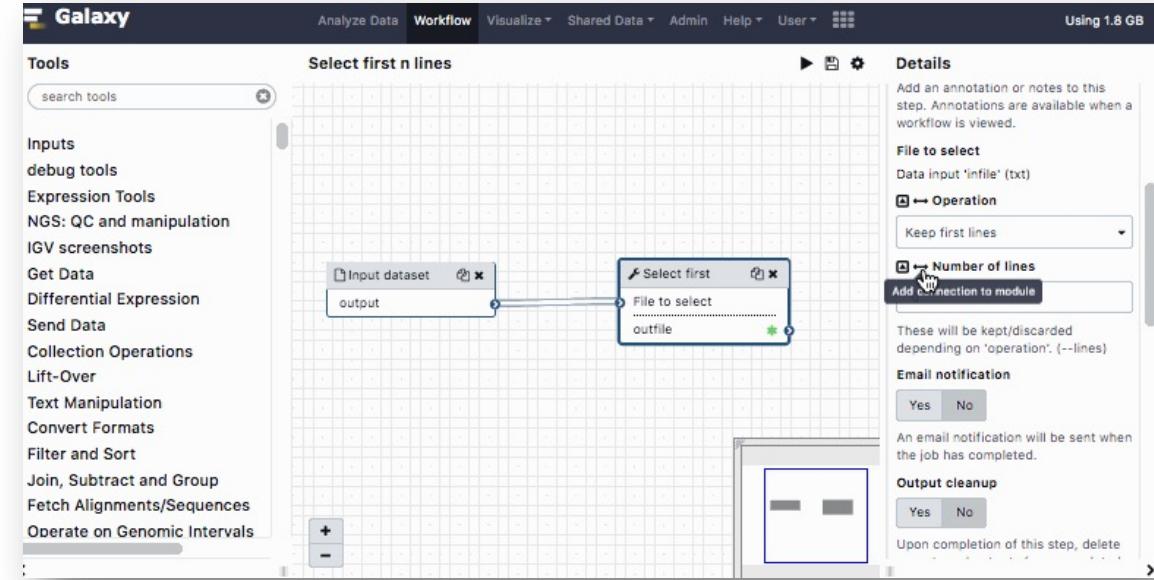
MetaPhlAn2, HUMAN2, VSearch, ...
QIIME, Mothur
MEGAHIT, MetaSpades, ...

Proteomics, Metabolomics, Chemistry

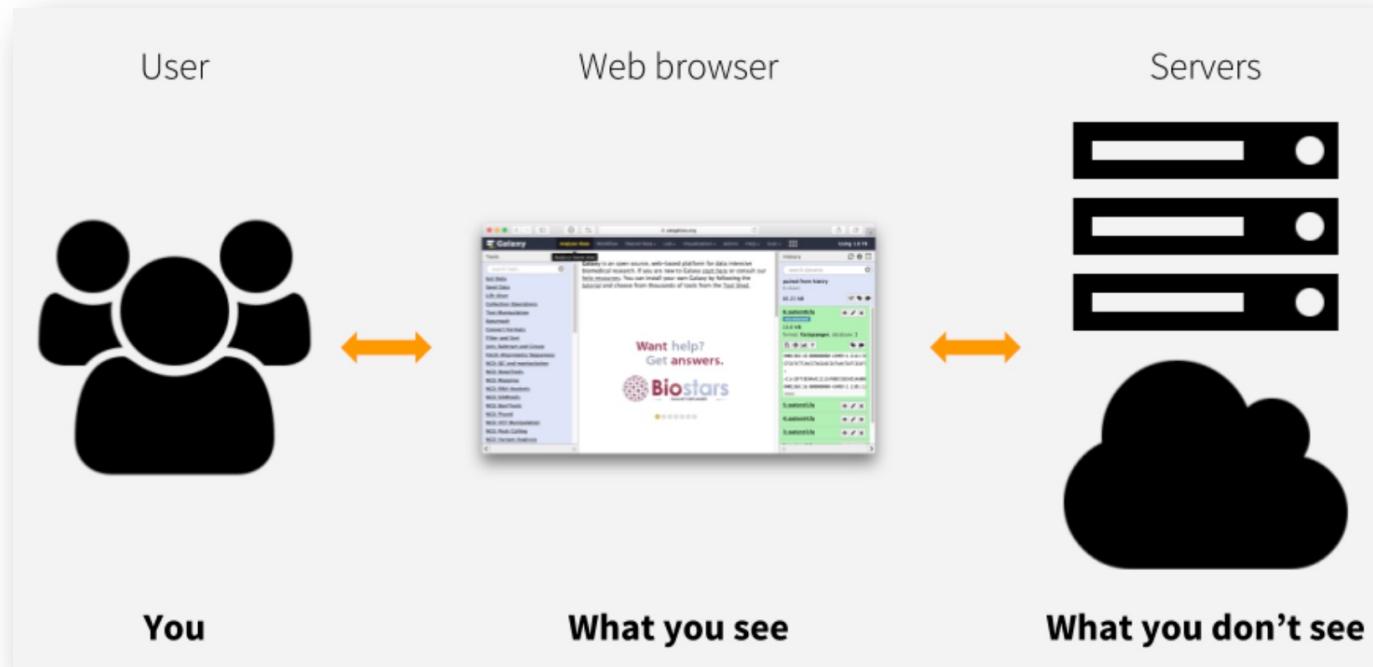
OpenMS, PeptideShaker, SearchGUI, MADLIquant, etc
Jmol Editor, Docking, etc
OpenBabel, ChemFP, OMG, QED, etc



- **Accessibility**
 - Users without programming experience can easily upload/retrieve data, run complex tools and workflows, and visualize data
- **Reproducibility**
 - Galaxy captures information so that any user can understand and repeat a complete computational analysis
- **Transparency**
 - Users can share or publish their analyses (histories, workflows, visualizations)
 - Pages: online Methods for your paper



Galaxy PROJECT



- 120+ public servers, many more non-public
- Both general-purpose and domain-specific
- Free, limitations apply: 250GB storage
- <https://usegalaxy.org>
- <https://usegalaxy.eu>

Tutorials and workflows Epigenetics

Material

Search x

Lesson	Slides	Hands-on	Input dataset	Workflows	Galaxy tour	Galaxy instances
Introduction to ChIP-Seq data analysis						
Introduction to DNA Methylation data analysis						
ATAC-Seq data analysis						
DNA Methylation data analysis						
Formation of the Super-Structures on the Inactive X ChIP-seq						
Hi-C analysis of Drosophila melanogaster cells using HiCExplorer						
Identification of the binding sites of the T-cell acute lymphocytic leukemia protein 1 (TAL1) ChIP-seq						
Infinium Human Methylation BeadChip						

Domain-specific language (DSL) workflow managers

- DSL is a programming language developed to meet a specific need in a particular domain
- WFM implemented in DSL
 - Reproducible, robust and portable
 - Incorporate existing tools
 - Reusable modules
- Maximal flexibility, but can have a bit of an initial learning curve
(GUIs being developed to improve accessibility)
- Nextflow, Snakemake, ...

Programming-library-based workflow managers

- Less common than the two previous WFM
- Pipeline management for an existing programming language
- SciPipe for Go, Luigi for Python, Targets for R, ...
- Minimizes learning curve if already familiar with language
- Fewer features, less adopted for bioinformatics application

No need to reinvent the wheel!

- Often good idea to look for existing pipelines
- Best-practice pipeline collections
- Galaxy published workflows, KNIME, nf-core, Snakemake-Workflows,
...

Pipeline initiative	Tool	Curated ^a	Community ^b	Citable ^c	Pipelines ^d
nf-core	Nextflow	✓	✓	✓	27
snakePipes	Snakemake	✓	✗	✓	9
Snakemake-Workflows	Snakemake	✓	✓	✓	7
GenPipes	GenPipes	✓	✗	✓	12
Galaxy Community	Galaxy	✗	✓	✗	>1,000
BioWDL	WDL	✓	✗	✓	17
WARP	WDL	✓	✗	✗	8
KNIME Hub	KNIME	✗	✓	✗	>1,000

Use a workflow manager!

Whenever you:

- Use more than 1 tool
- Your analysis contains long calculation steps
- Expect to repeat your analysis
- Expect to publish or share your analysis results
- Want to avoid tricky software installation