GHGA

Webinar

Introduction to the Standardization of NGS Workflows

30.05.2023
16:00 (CEST)

Kübra Narcı
DKFZ

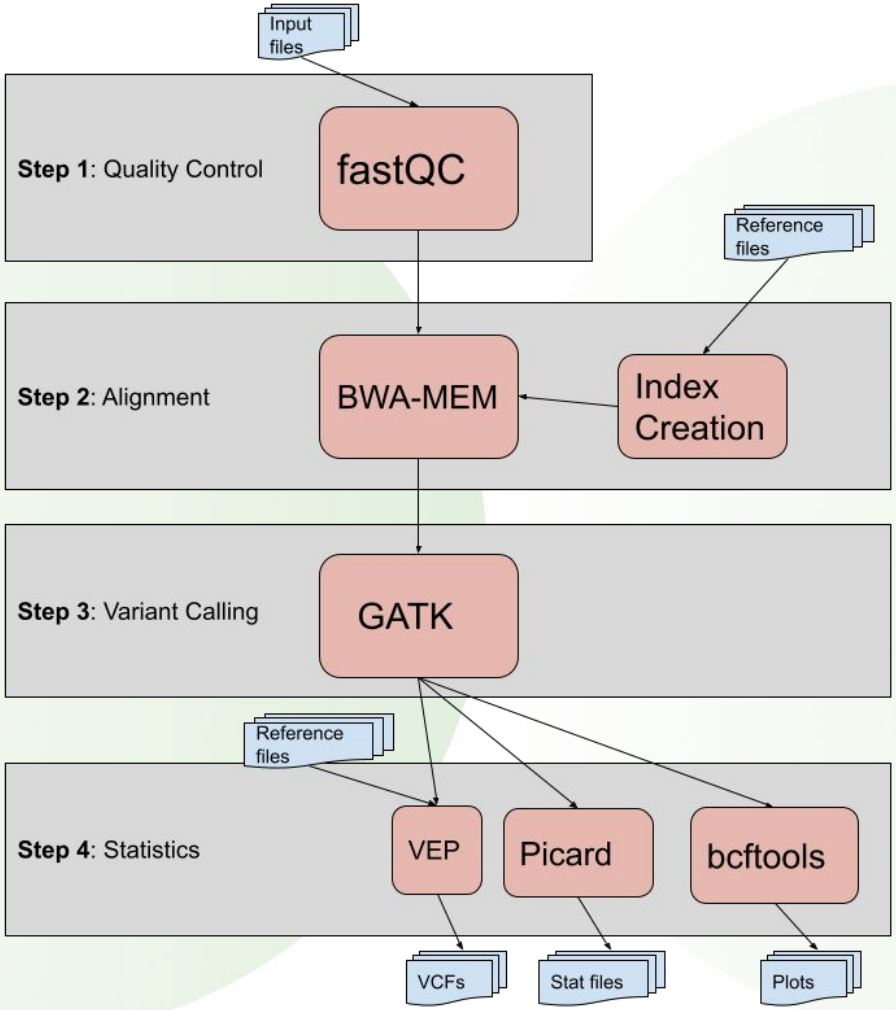
Outline

- Traditional bioinformatic workflows
  - Challenges
  - Reproducibility crisis
- GHGA goals and objectives
- How to standardize workflows
  - Components
- Workflow management systems
- Summary

GHGA

Bioinformatics Workflows

- ‘Software tool-chains’
- Purpose: Automation
- Traditional forms are problematic:
  - Adaptability
  - Various infrastructure needs
  - Fragile ecosystems



GHGA

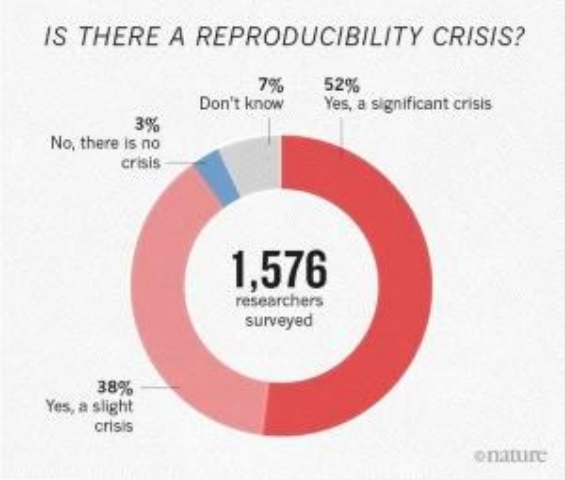
Common Challenges

- Reproducing results
- Software dependencies
- Software updates
- Sharing workflows
- Scalable analysis

GHGA

Reproducibility Crisis

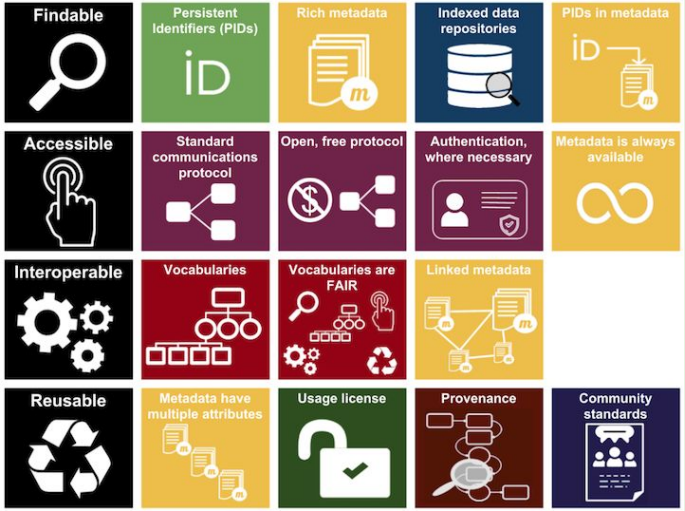
- **Science depends on researchers being able to replicate the work of others**



Baker, M. 1,500 scientists lift the lid on reproducibility. Nature 533, 452–454 (2016). https://doi.org/10.1038/533452a

Reproducibility Crisis

- Science depends on researchers being able to replicate the work of others

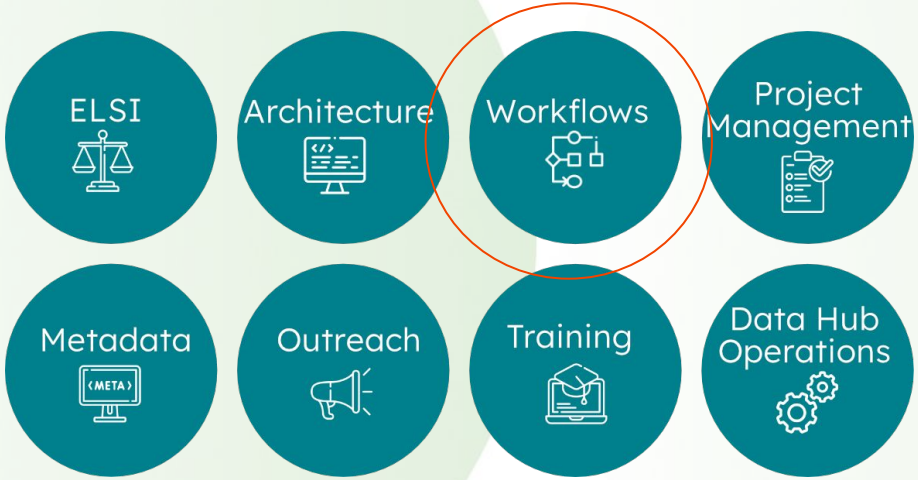


FAIR
https://doi.org/10.1038/sdata.2016.18

GHGA

GHGA Goals & Objectives

- Establish a national infrastructure for human omics data
- An ethico-legal and data use framework for
  - Data Sharing
  - Protection
  - Analysis
- Provide standards for:
  - Metadata and Workflows
- Make human omics data FAIR
  - being GA4GH compliant



Global Alliance for Genomics & Health

Collaborate. Innovate. Accelerate.

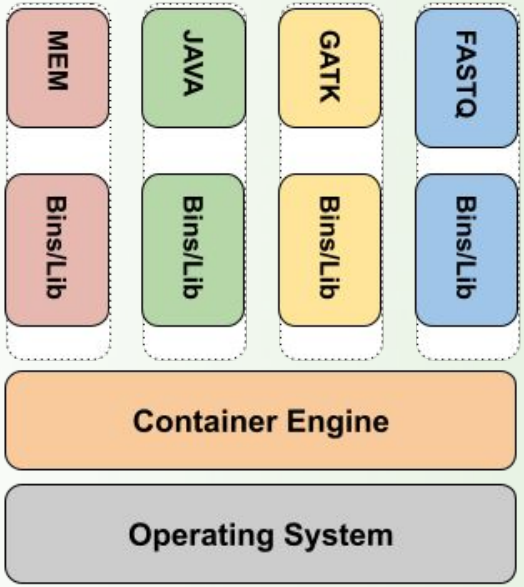
GHGA

Learning Goals

- Components and principles of standardized bioinformatic workflows
- Metrics to measure the FAIRness of a workflow
- Processes to automate workflows
- Workflow management systems
- Recommendations for workflow developers

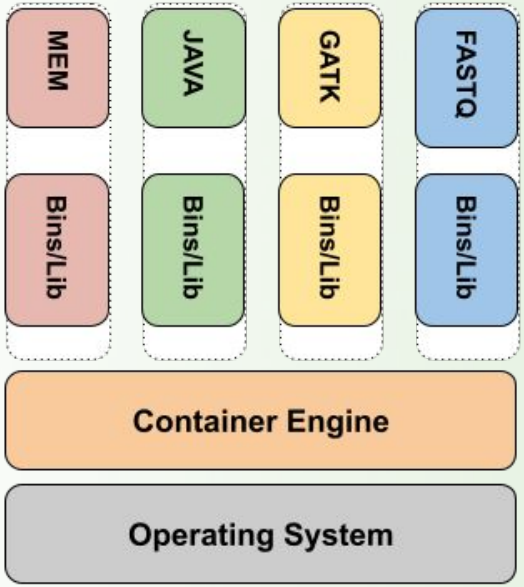
GHGA

Portability



- Automation of installation
- Configuration

Portability

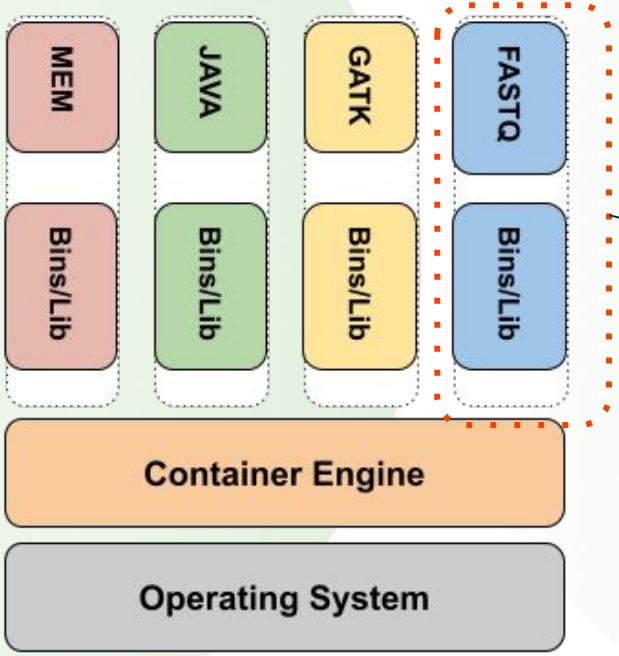


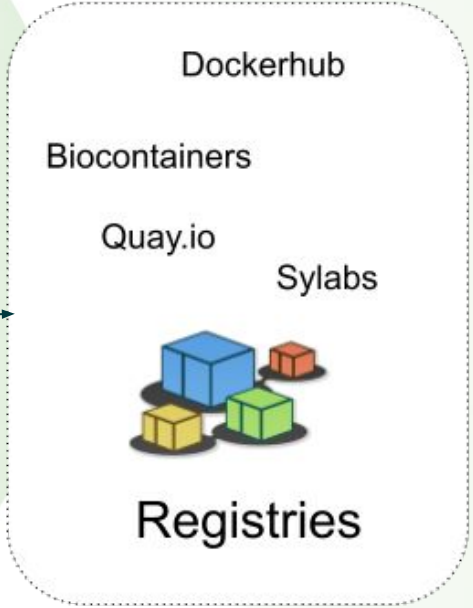
- Packaging and distribution
- Platform-independent
- Modular

GHGA

Portability

Containers are separate, but bins/libs and OS are shared.





GHGA

Versioning

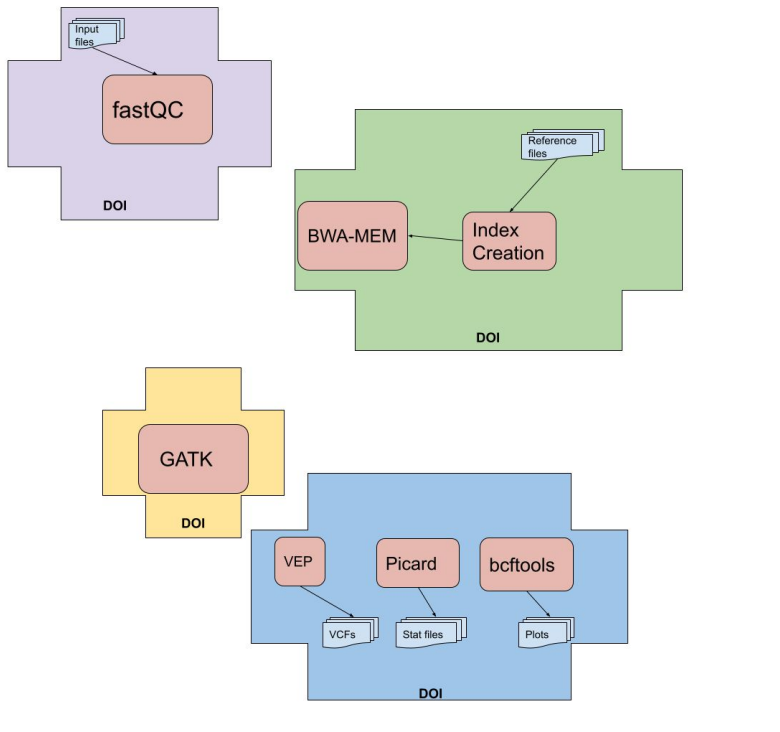
Readme | Packages and Containers | Similar Tools

Type	Version	Last Update	Size	Full Tag	Security
docker	v3.14.0-1-deb	2018-04-04	71.21M	docker pull biocontainers/fastaq:v3.14.0-1-deb_cv1	
docker	v3.17.0-2-deb	2019-09-12	244.46M	docker pull biocontainers/fastaq:v3.17.0-2-deb_cv1	

- Versioning and registering containers
- Repurposing images and sharing own repositories

Findability

- Rich metadata
- Unique and persistent identifier (DOI)
- Searchable sources



zenodo

Dockstore

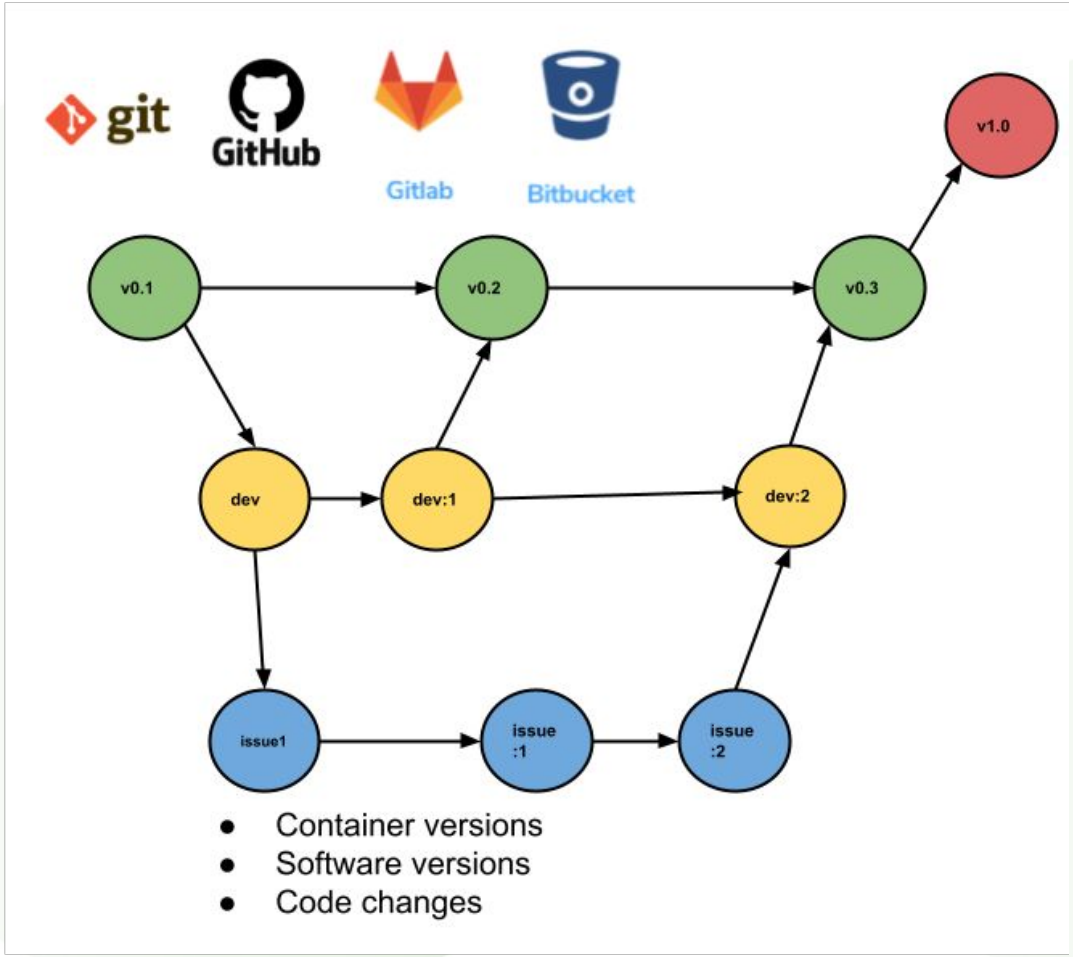
nf-core

Galaxy

GHGA

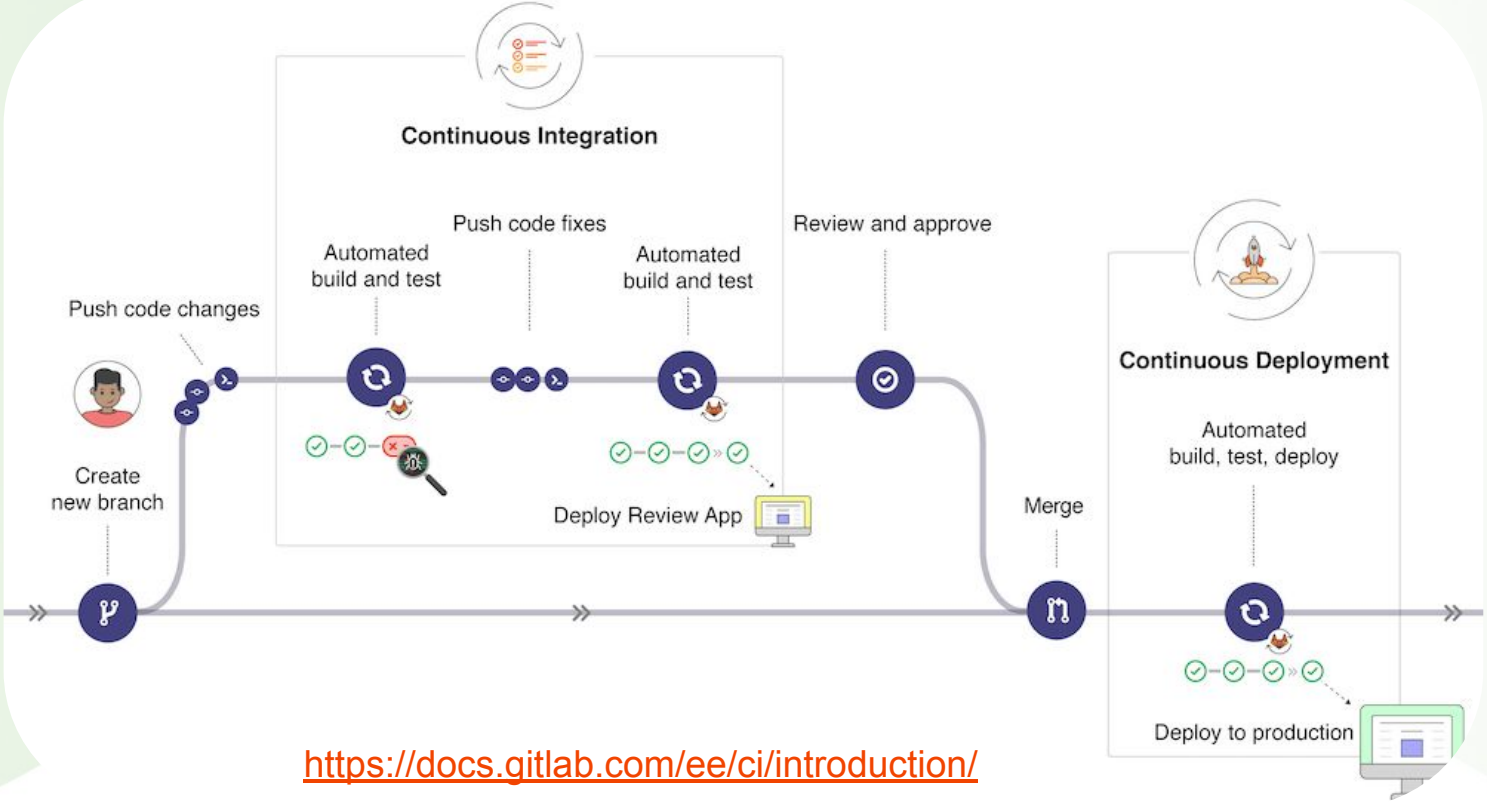
Stability

- Storage
- Collaborating
- Authorized control



GHGA

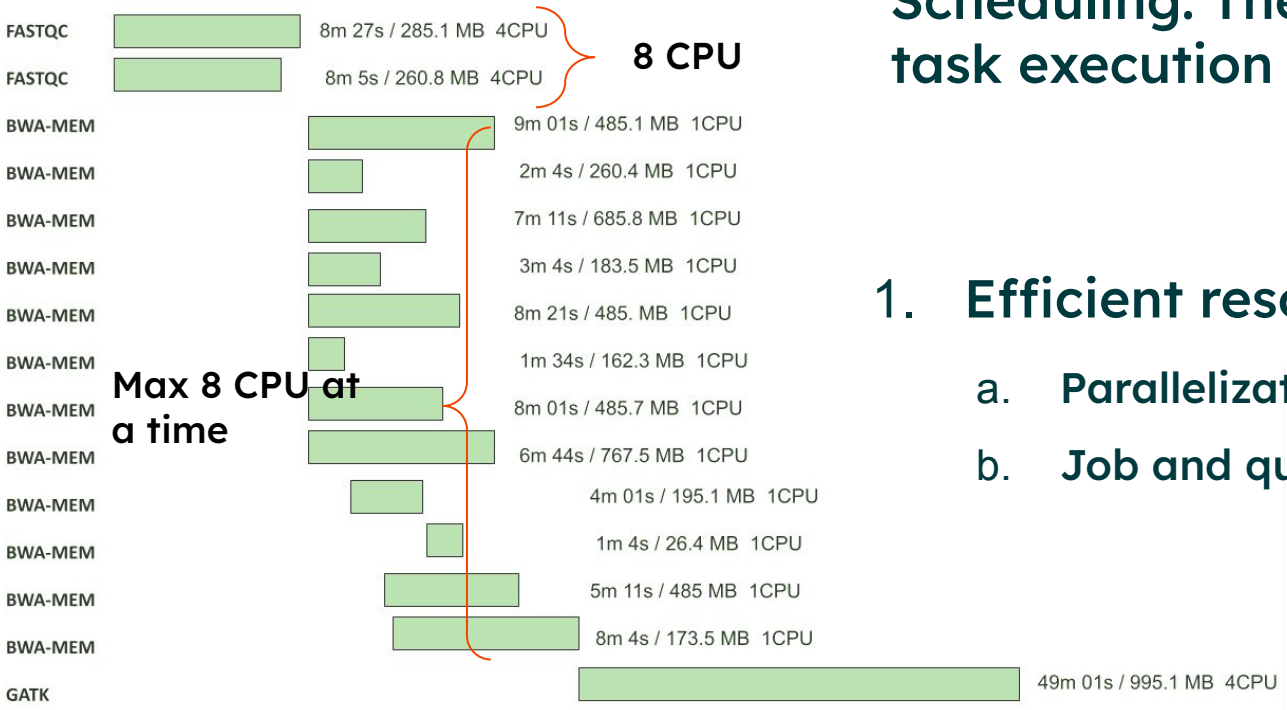
Stability



https://docs.gitlab.com/ee/ci/introduction/

GHGA

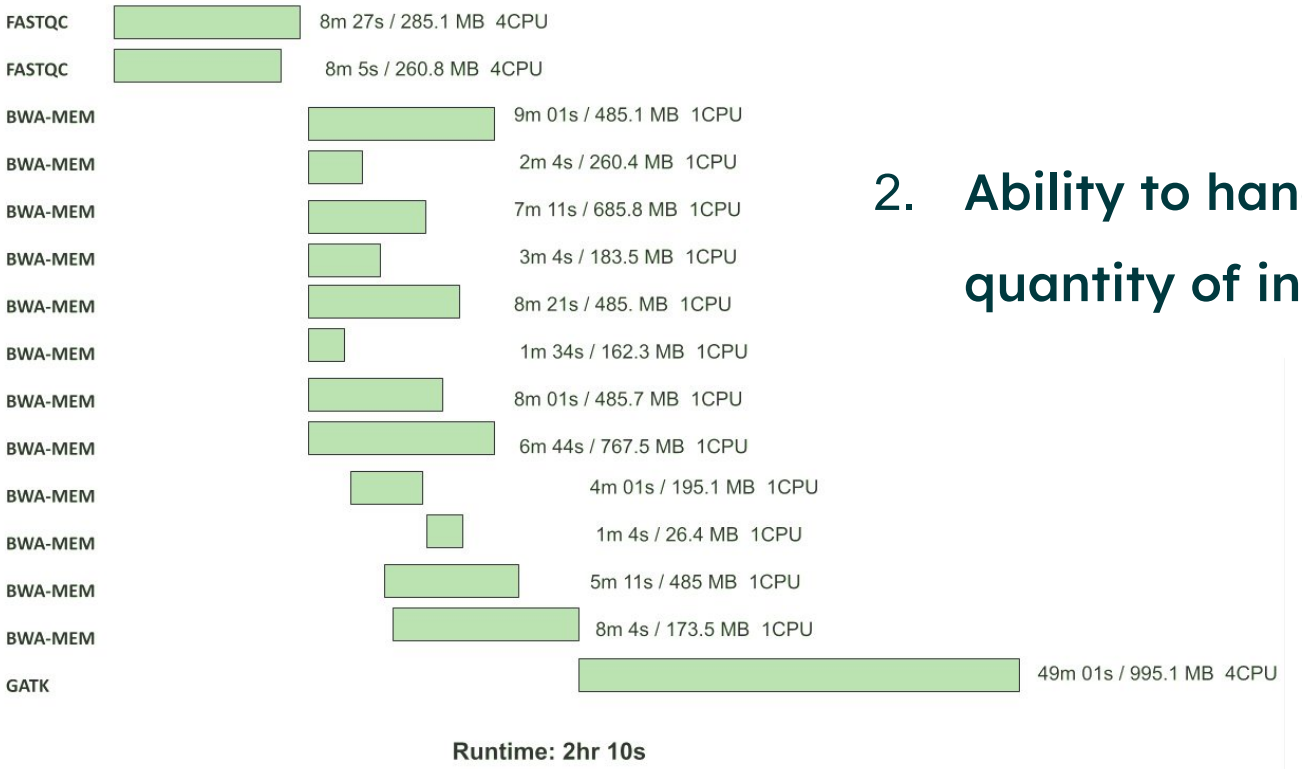
Scalability



Scheduling: The art of controlling task execution

1. **Efficient resource management**
   a. **Parallelization of different steps**
   b. **Job and queue scheduling**

Scalability



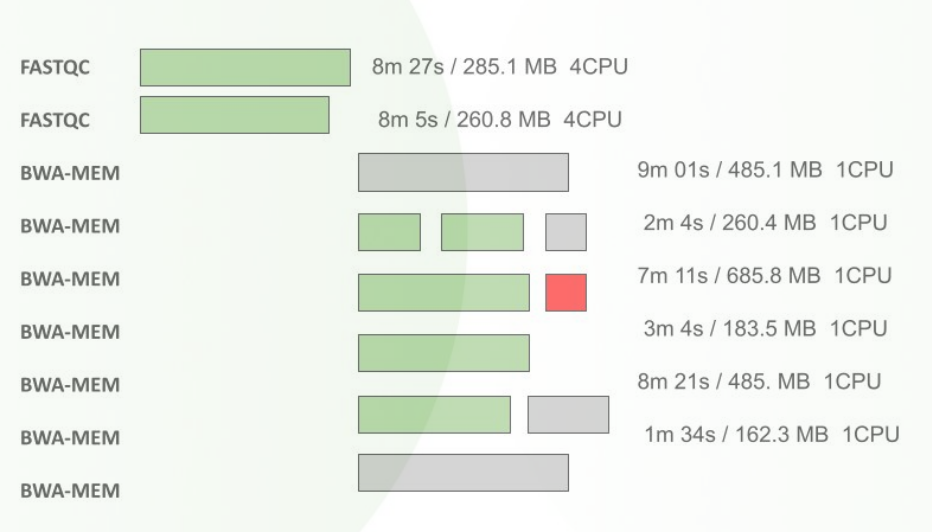
2. **Ability to handle any size and quantity of input data**

AWS BATCH

slurm workload manager

kubernetes

Re-entrancy



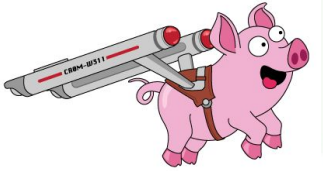
- Caching intermediate results
- Avoiding waste of
  - cost
  - recalculations
  - data processing
  - time

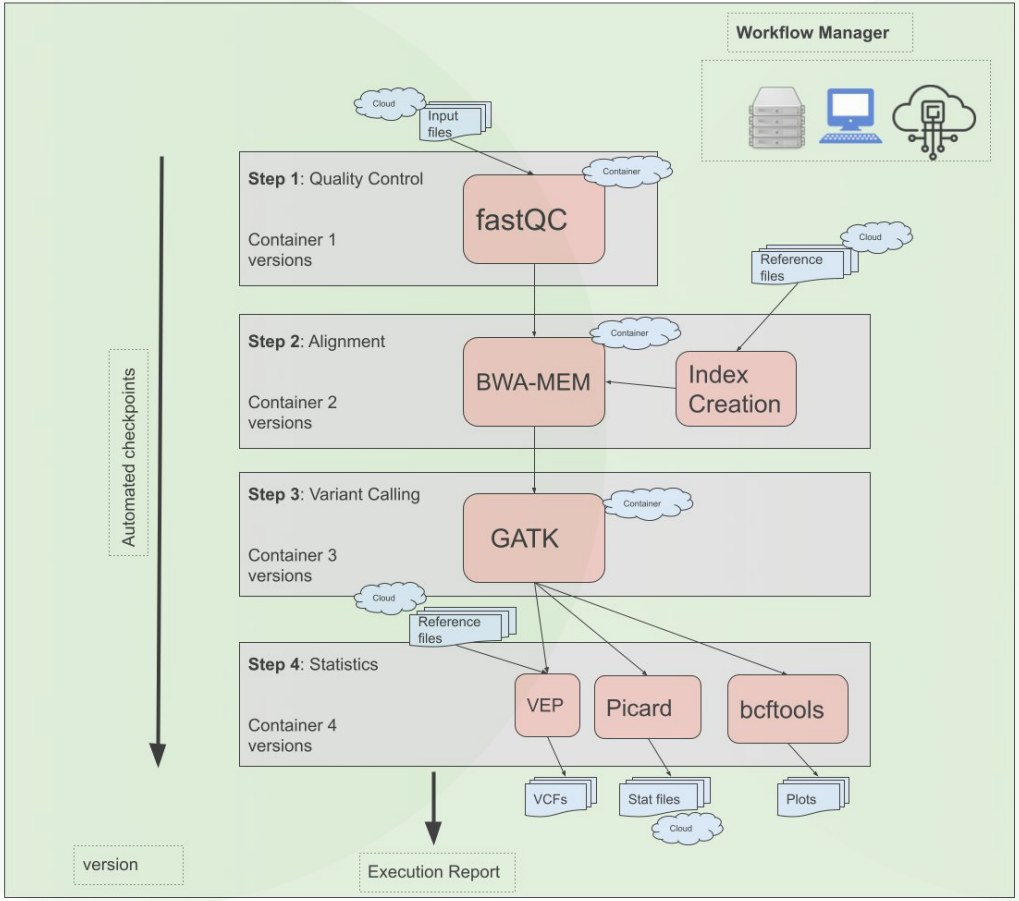
GHGA

QA

GHGA

Workflow Management Systems





GHGA

Provenance

The history of a computational experiment

Prospective

- Specifications of workflow
- Workflow steps
- Execution order
- Expected inputs/outputs
- Versions of softwares
- Versions of data

Retrospective

- Information about execution
- Produced inputs/outputs
- Used data
- Artifacts
- Consumed memory/cpu
- Process time

GHGA

Mondelli, M. L., Townsend Peterson, A., & Gadelha, L. M. R. (2019). Exploring Reproducibility and FAIR Principles in Data Science Using Ecological Niche Modeling as a Case Study. In Advances in Conceptual Modeling. ER 2019. Lecture Notes in Computer Science, vol. 11787 (pp. 23–33). Springer. https://doi.org/10.1007/978-3-030-34146-6_3

Provenance

The history of a computational experiment

Prospective

- Specifications of workflow
- Workflow steps
- Execution order
- Expected inputs/outputs
- Versions of softwares
- Versions of data

Retrospective

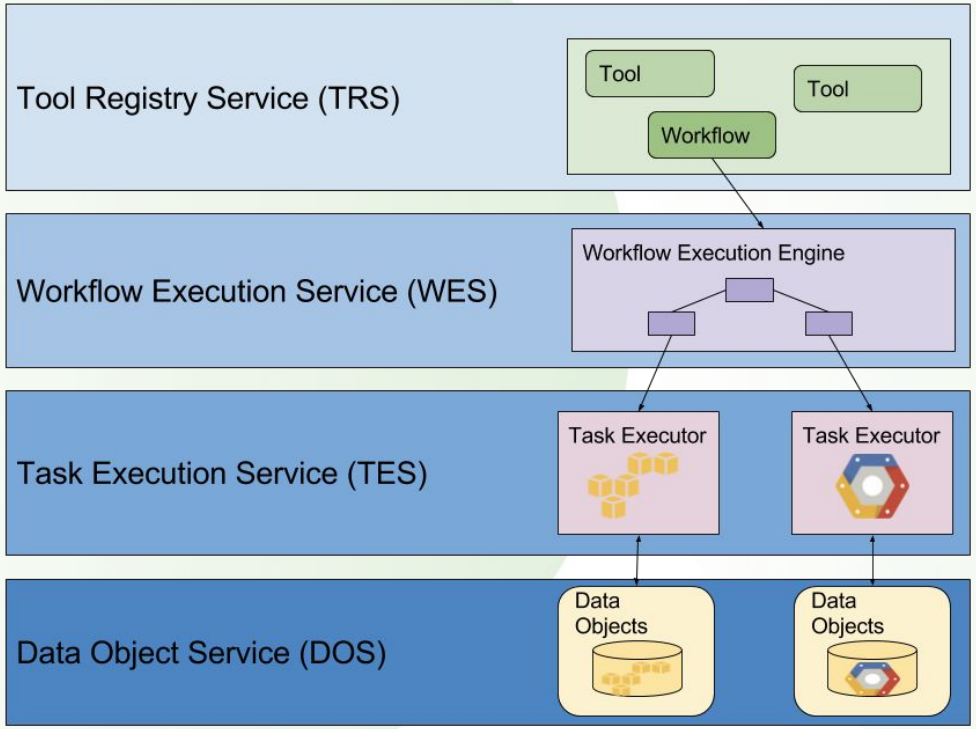
- Information about execution
- Produced inputs/outputs
- Used data
- Artifacts
- Consumed memory/cpu
- Process time

GHGA

Mondelli, M. L., Townsend Peterson, A., & Gadelha, L. M. R. (2019). Exploring Reproducibility and FAIR Principles in Data Science Using Ecological Niche Modeling as a Case Study. In Advances in Conceptual Modeling. ER 2019. Lecture Notes in Computer Science, vol. 11787 (pp. 23–33). Springer. https://doi.org/10.1007/978-3-030-34146-6_3

Computing Environment

- Sharing tools and workflows
- Executing workflows
- Executing individual tasks
- Accessing data

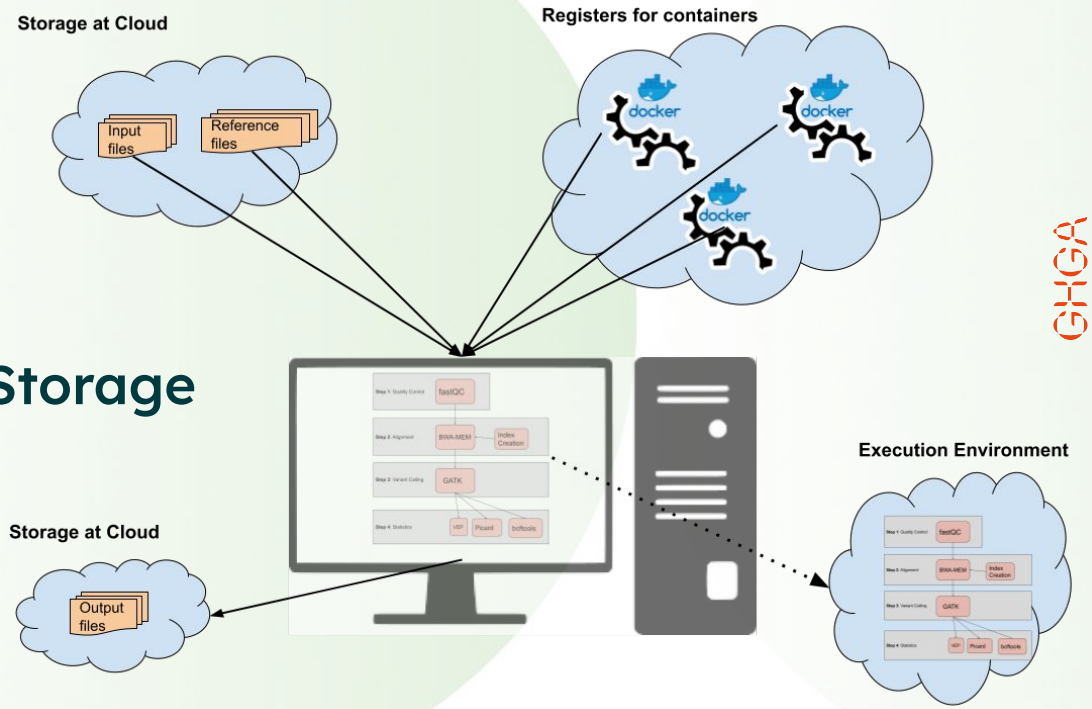


https://ga4gh-cloud.github.io/

GHGA

Accessing Data

- Data Repositories
  - iGenomes (Illumina)
  - Refgenie
  - Galaxy
- Database, Compute and Storage
  - AWS S3
  - Google Bucket
  - Azura Bucket



Quality Control & Management

1 **Quality Control**
- Implementation of QC inside the workflow

2 **Post Processing**
- Cleaning / Normalisation
- Checking statistics

3 **Optimization**
- Parameters
- Speed & Accuracy

4 **Validation**
- Benchmarking
- Speed & Accuracy

5 **Provenance**
- History of runs

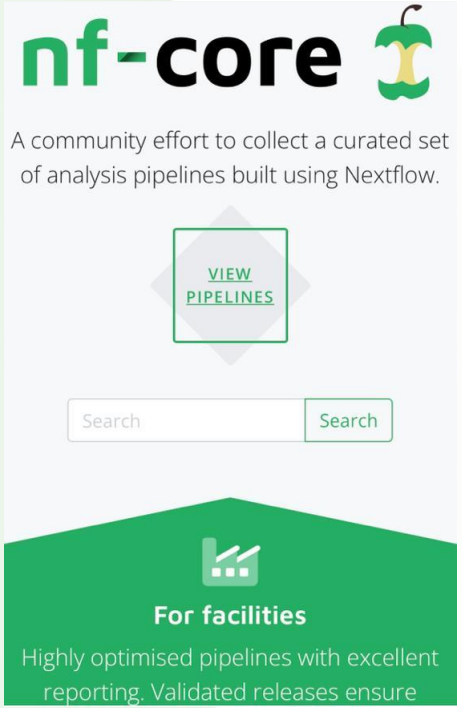
GHGA

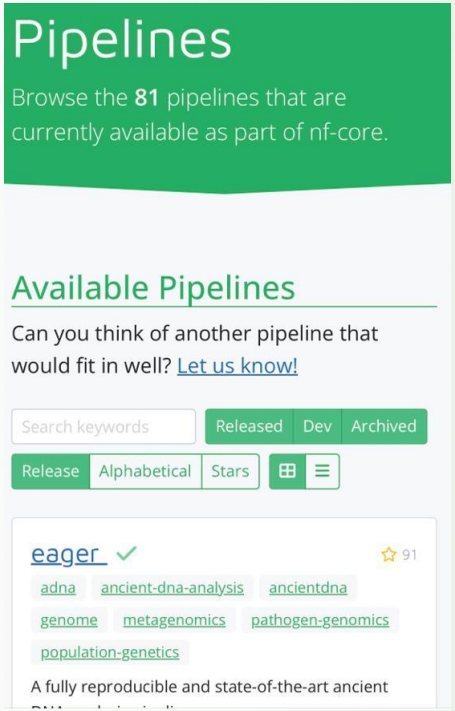
Increasing Productivity

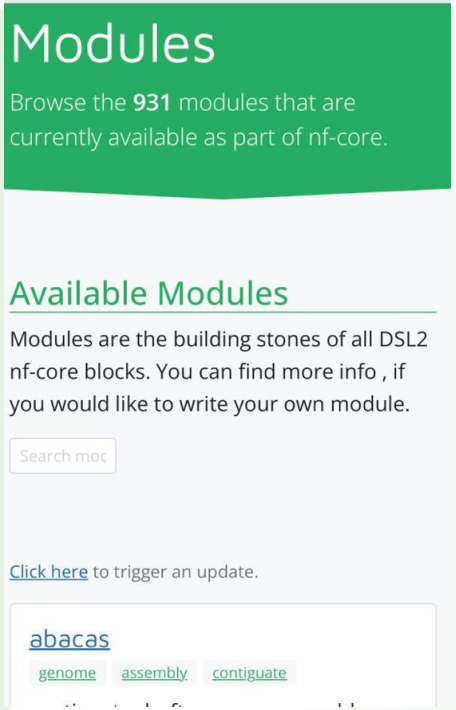
- Find your favorite
  - Code editor
  - Text editor
- Run and debug
- Documentation
- Task management

GHGA

Community Driven Bioinformatic Workflows







GHGA

What does nf-core provides for GHGA ?

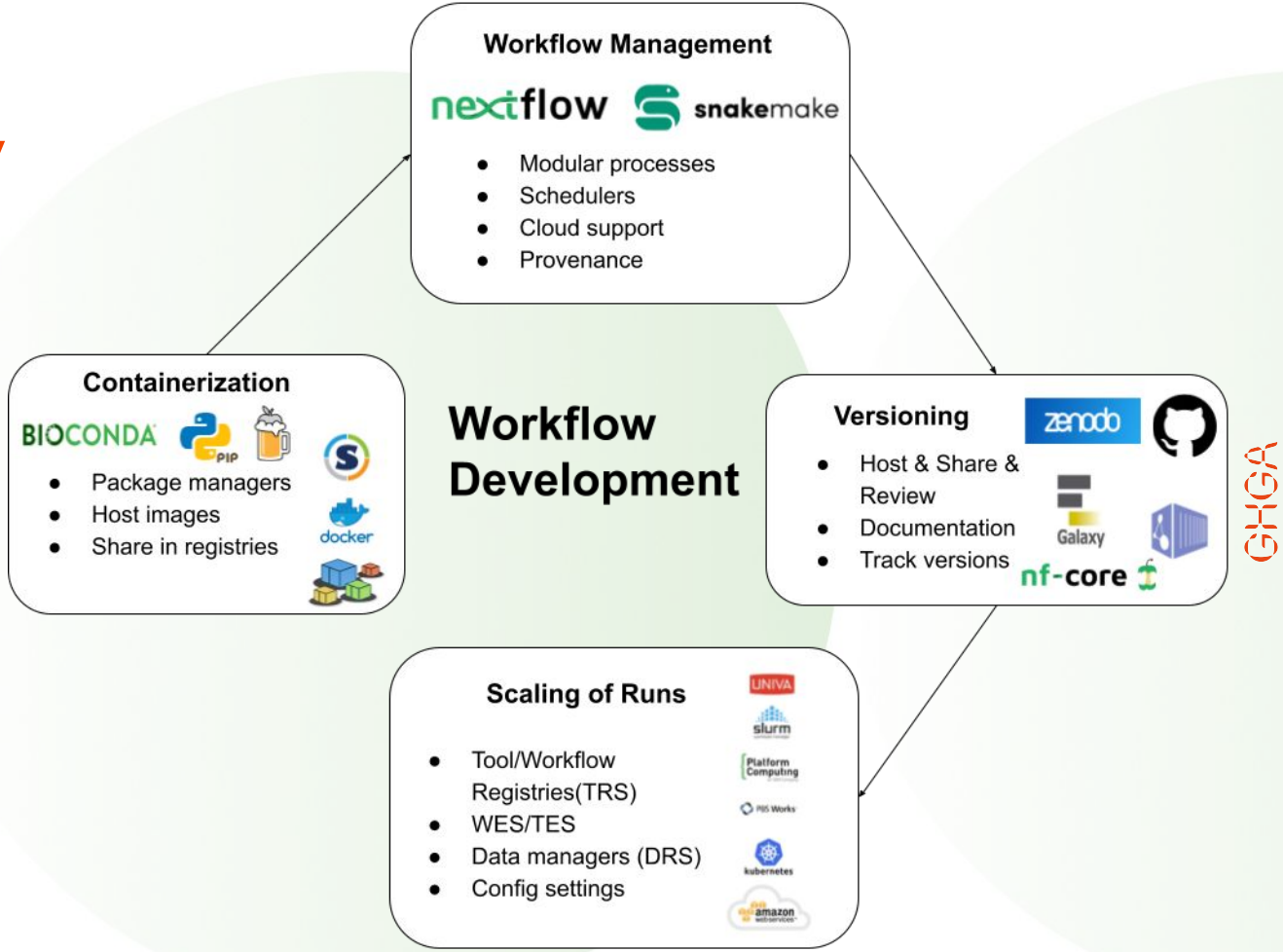
- Documentation
- CI Testing
- Stable Releases
- Packaged software
- Portable and reproducible
- Cloud-ready

https://github.com/nf-core

GHGA

Summary

- Accurate
- Scalable
- Reproducible
- Portable



https://github.com/ghga-de

Webinar

Introduction to Benchmarking of NGS Workflows

20.06.2023
16:00 (CEST)

Kübra Narcı
DKFZ

Thank you!

And thanks to the team and the PIs!

Florian Heyl

Paul Menges

Kübra Narci

Luiz Gadelha

Evangelos Theodorakis

Christian Mertes

Resources

- Nextflow sources: https://www.nextflow.io/
- Nf-core: https://nf-co.re/
- GH4GA: https://www.ga4gh.org/
- Writing and sharing portable tools and workflows
  - https://dockstore.org
  - https://biocontainers.pro
- Papers:
  - Sandve GK, Nekrutenko A, Taylor J, Hovig E (2013) Ten Simple Rules for Reproducible Computational Research. PLoS Comput Biol 9(10): e1003285. https://doi.org/10.1371/journal.pcbi.1003285
  - Spjuth O, Capuccini M, Carone M et al. Approaches for containerized scientific workflows in cloud environments with applications in life science [version 1; peer review: 2 not approved]. F1000Research 2021, 10:513 (https://doi.org/10.12688/f1000research.53698.1)
  - Carole Goble, Sarah Cohen-Boulakia, Stian Soiland-Reyes, Daniel Garijo, Yolanda Gil, Michael R. Crusoe, Kristian Peters, Daniel Schober; FAIR Computational Workflows. Data Intelligence 2020; 2 (1-2): 108–121. doi: https://doi.org/10.1162/dint_a_00033

GHGA