seqeralabs

Marcel Ribeiro-Dantas, PhD
mribeirodantas@seqera.io

Reproducibility for Data Pipelines and Analyses

Nextflow & nf-core

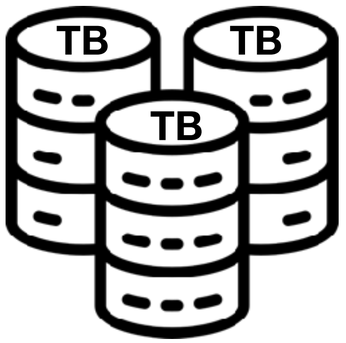
Open Science

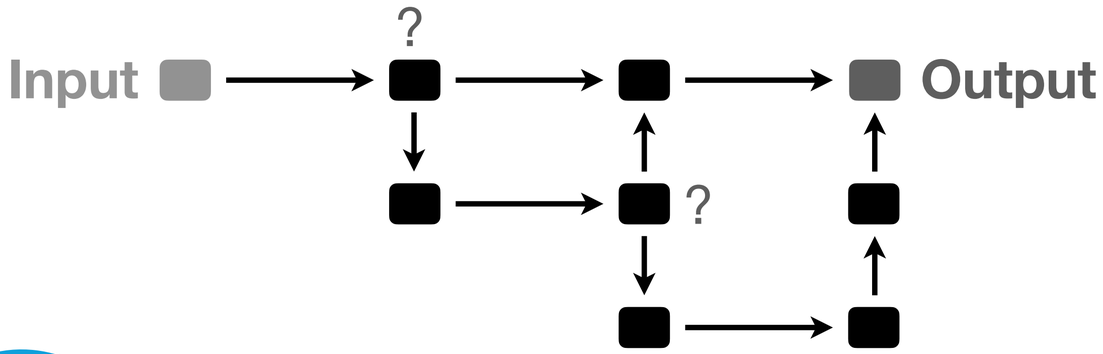
Open Source

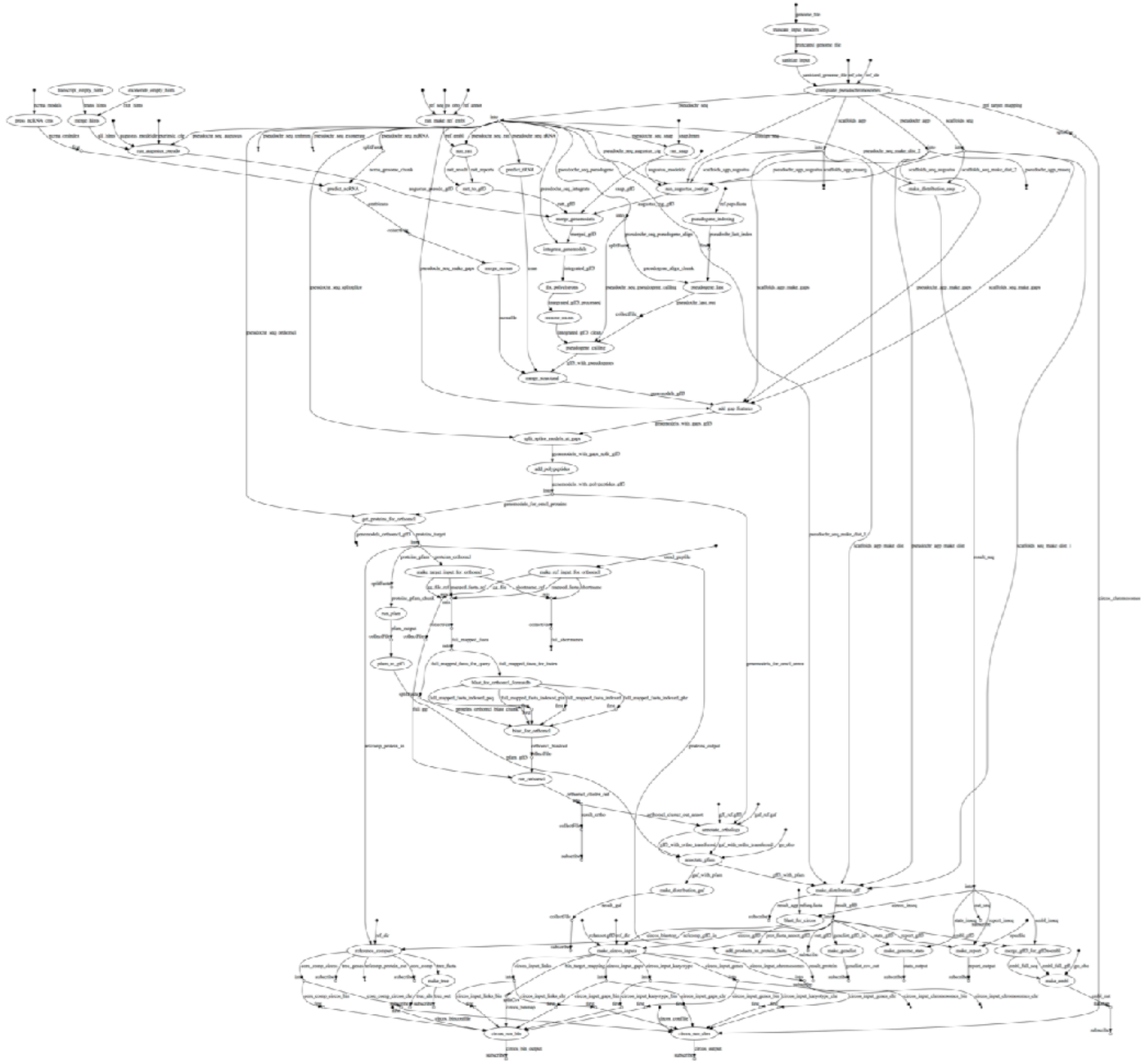
Open Data

Open Data Community

Genomics Workflow







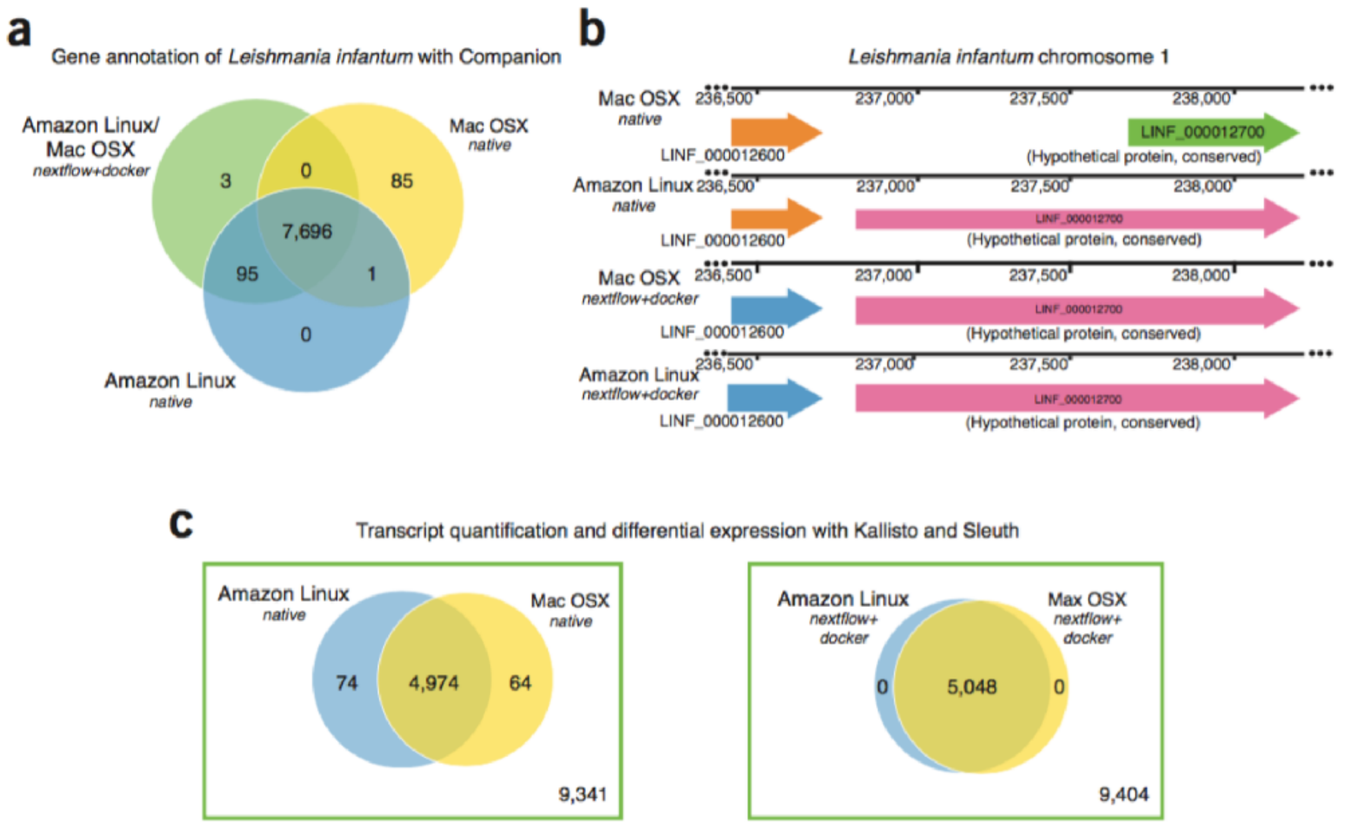
70 tasks
55 custom scripts
39 softwares tools & libraries

Steinbiss et al., Companion parassite genome annotation pipeline
DOI: 10.1093/nar/gkw292

Nextflow enables reproducible computational workflows

Paolo Di Tommaso, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo & Cedric Notredame

NATURE BIOTECHNOLOGY VOLUME 35 NUMBER 4 APRIL 2017

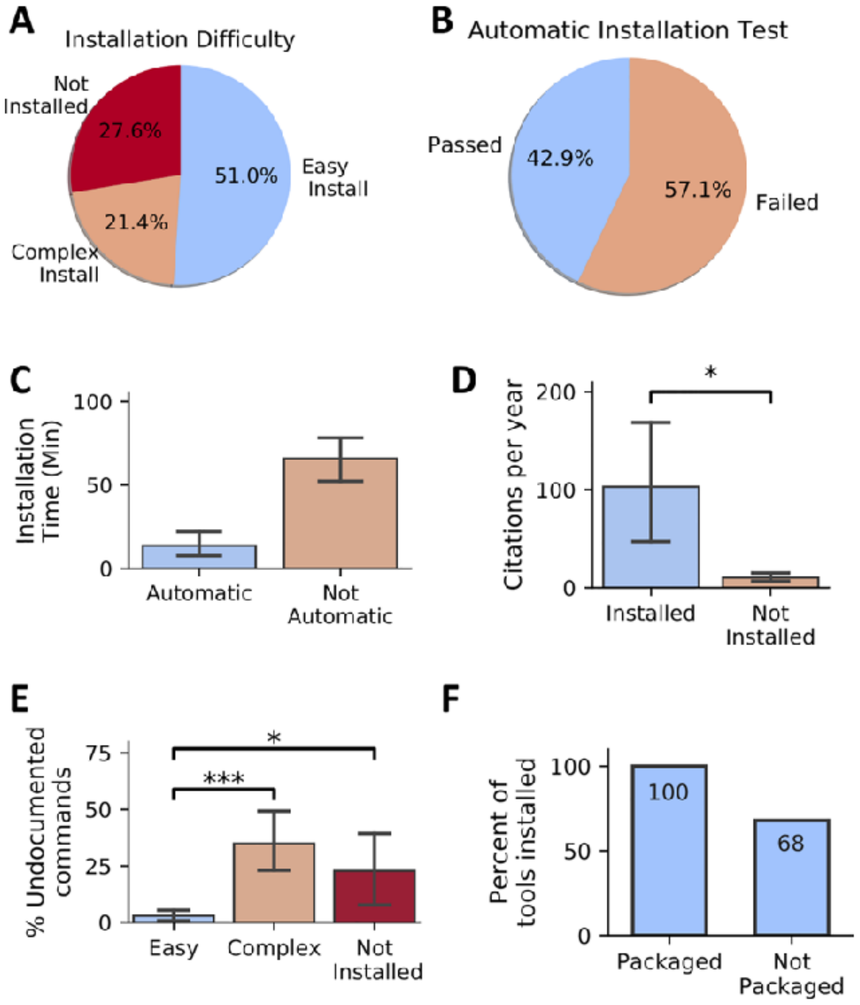


Challenges and recommendations to improve the installability and archival stability of omics computational tools (2019)

Serghei Mangul, et al. Plos Biology https://doi.org/10.1371/journal.pbio.3000333

We found that 28% of all omics software resources are currently not accessible through URLs published in the paper.

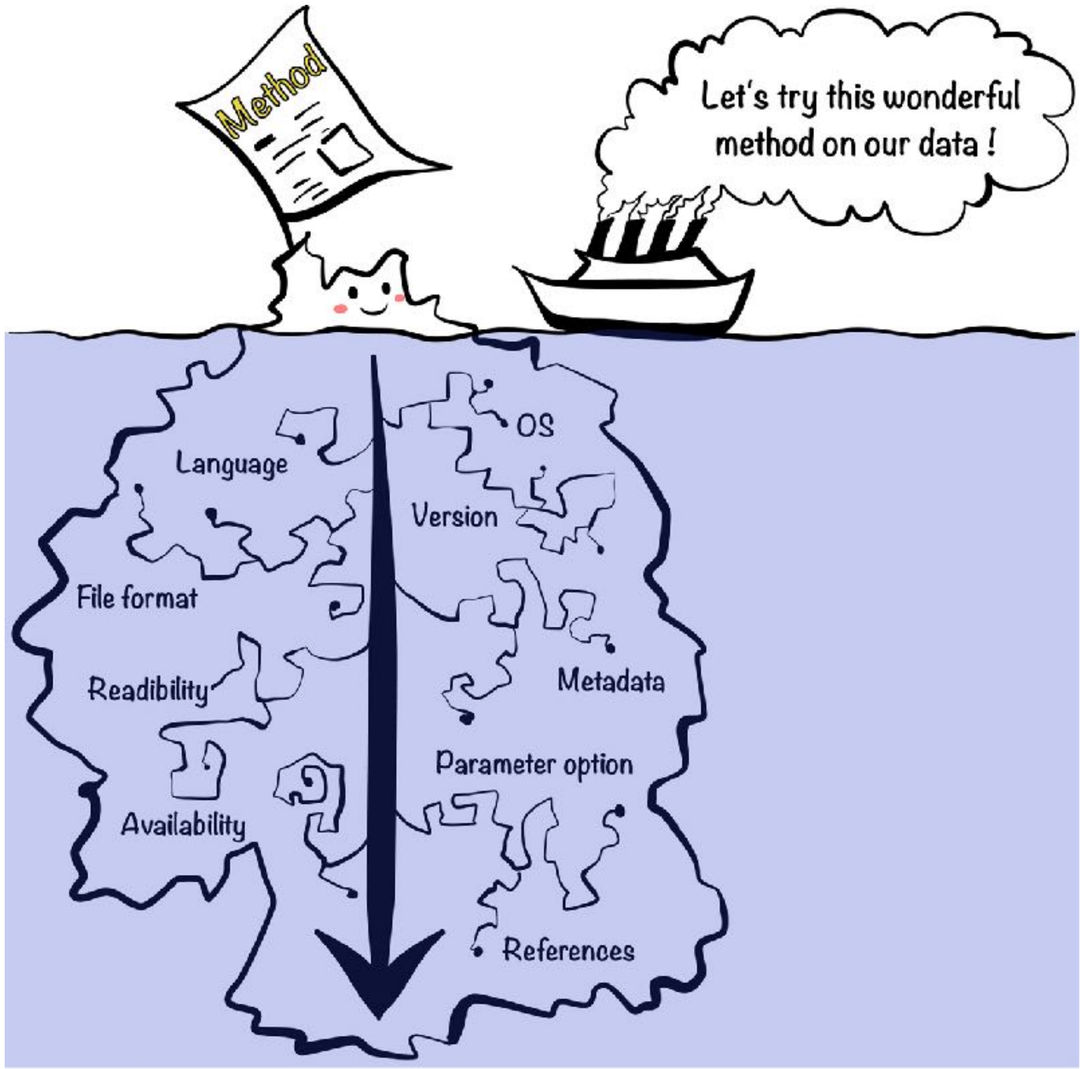
Among the tools selected 49% were difficult to install or could not be installed at all.



Comparison of the Companion pipeline annotation of Leishmania infantum genome executed across different platforms *

Platform	Amazon Linux	Debian Linux	Mac OSX
Number of chromosomes	36	36	36
Overall length (bp)	32.032.223	32.032.223	32.032.223
Number of genes	7.781	7.783	7.771
Gene density	236,64	236,64	236,32
Number of coding genes	7.580	7.580	7570
Average coding length (bp)	1.764	1.764	1.762
Number of genes with multiple CDS	113	113	111
Number of genes with known function	4.147	4.147	4.142
Number of t-RNAs	88	90	88

* Di Tommaso P, et al., Nextflow enables computational reproducibility, Nature Biotech, 2017



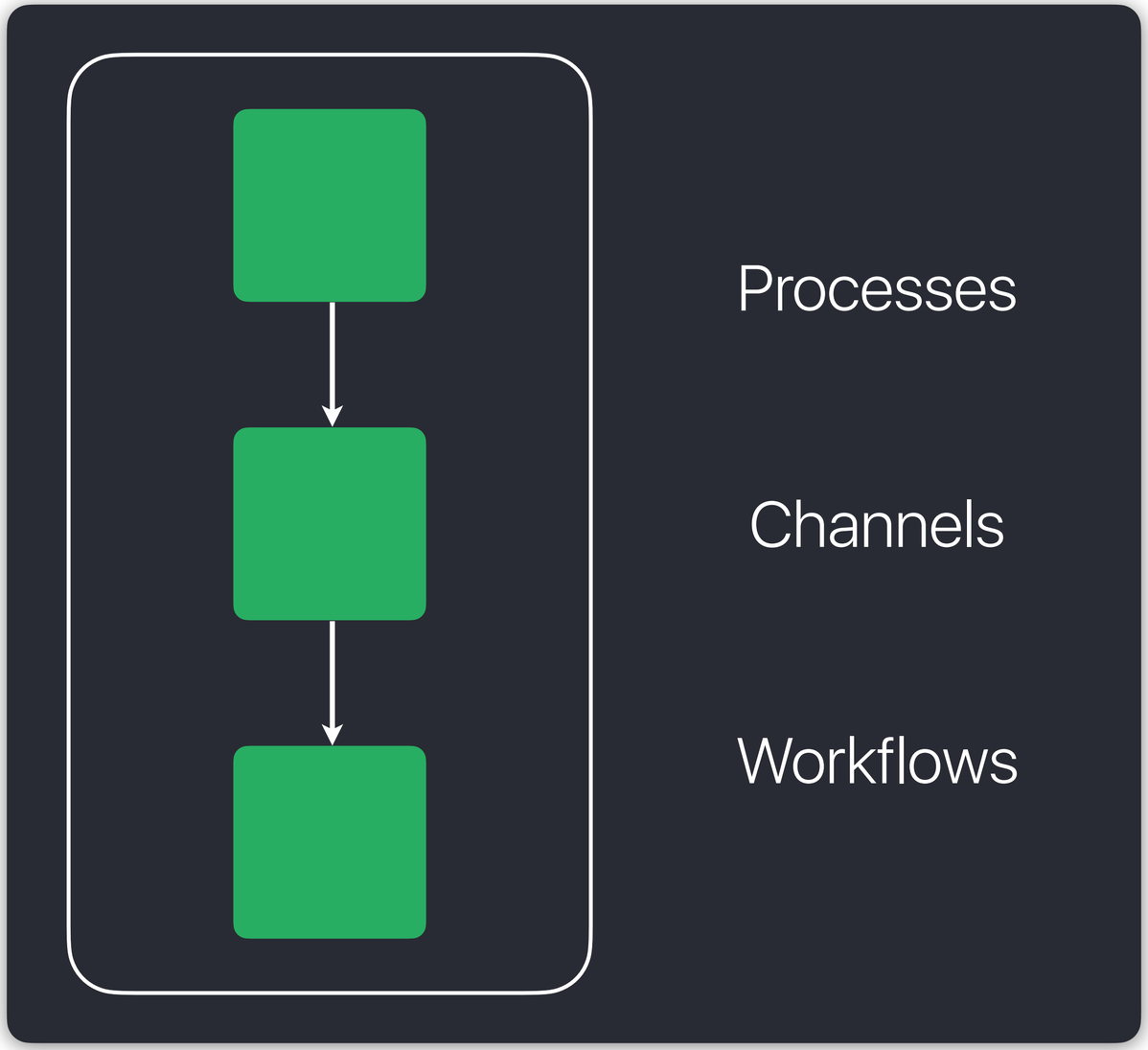
Kim et al. Experimenting with reproducibility: a case study of robustness in bioinformatics, *GigaScience*, Volume 7, Issue 7, July 2018.https://doi.org/10.1093/gigascience/giy077

nextflow

nextflow

Language

nextflow



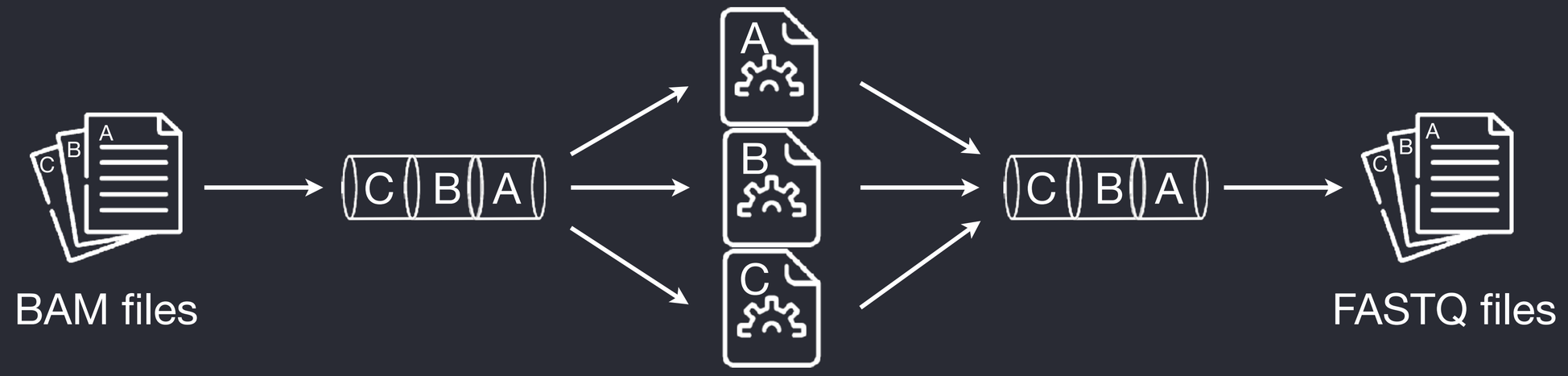
nextflow

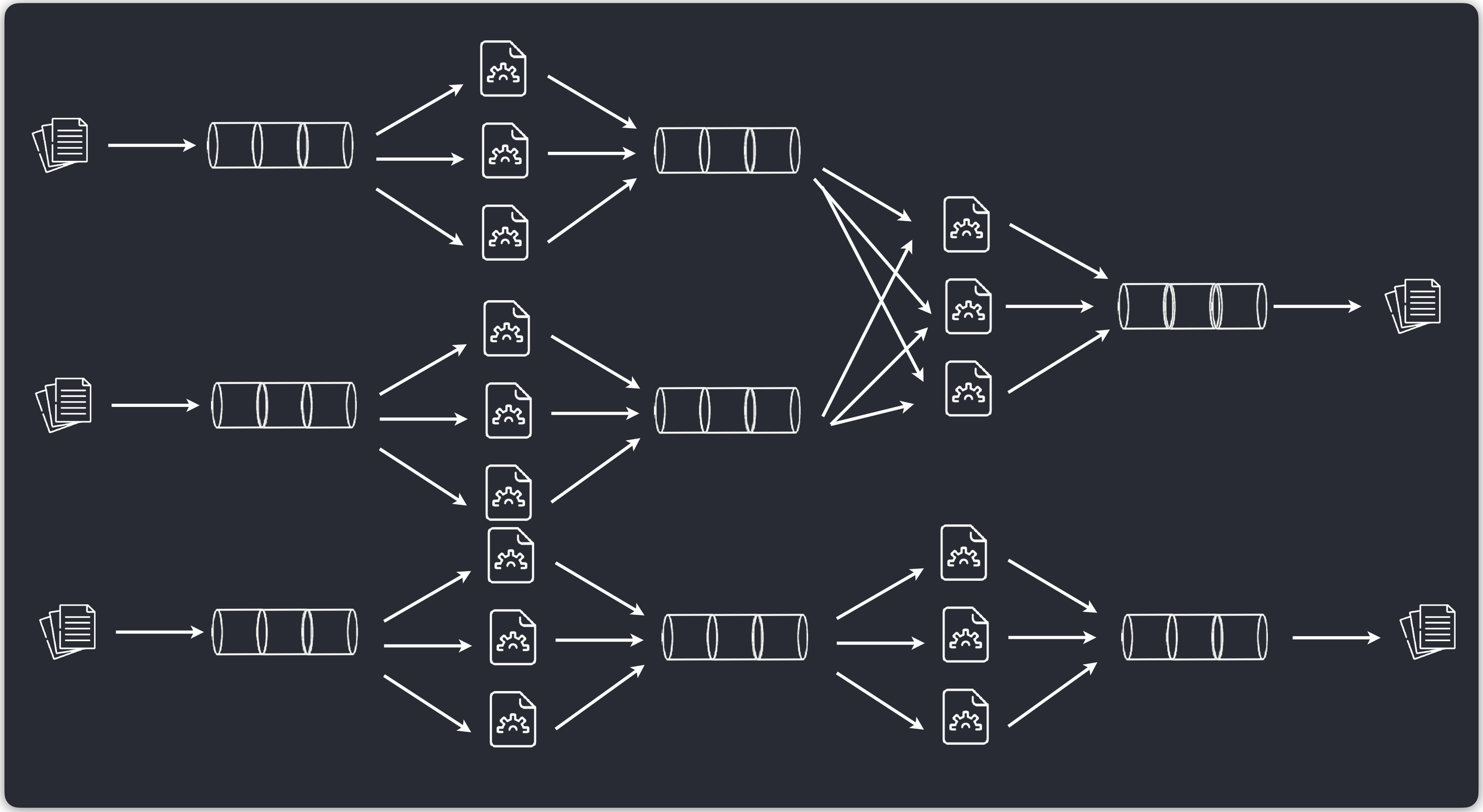
```
#!/usr/bin/env nextflow
process fastqc {
  input:
  path input

  output:
  path "*_fastqc.{zip,html}"

  script:
  """
  fastqc -q $input
  """
}

workflow {
  Channel.fromPath("*.fastq.gz") | fastqc
}
```

nextflow

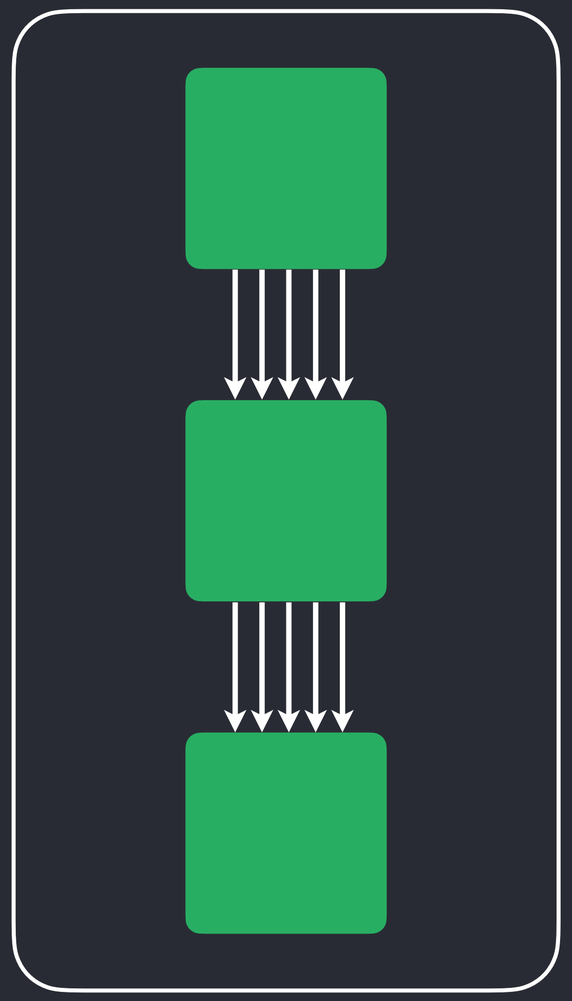
Implicit parallelism

```
workflow {
  Channel
    .fromPath("data/*.bam") | bam_to_fastq
}
```





nextflow



Parallelization

Reentrancy

(resume partial runs)

Reusability

nextflow

Language

Software

Compute

nextflow

git

GitHub

Bitbucket

GitLab

Gitea

AWS CodeCommit

Azure Repos

Software

Compute

nextflow

git

GitHub

Bitbucket

GitLab

Gitea

AWS CodeCommit

Azure Repos

docker

podman

Singularity

CONDA

Compute

nextflow

git

GitHub

Bitbucket

GitLab

Gitea

AWS CodeCommit

Azure Repos

docker

podman

Singularity

CONDA

SGE

openstack

Microsoft Azure

slurm
workload manager

aws

LSF

PBS

Google Cloud

kubernetes

nextflow

Reproducibility

```
process INDEX {

  input:
    path transcriptome
  output:
    path 'salmon_index'

  script:
  """
  salmon index --threads $task.cpus -t $transcriptome -i salmon_index
  """
}
```

nextflow

Reproducibility

```
process INDEX {
  conda "bioconda::salmon=1.9.0"
  input:
    path transcriptome
  output:
    path 'salmon_index'

  script:
  """
  salmon index --threads $task.cpus -t $transcriptome -i salmon_index
  """
}
```

nextflow

Reproducibility

```
process INDEX {
  container "nextflow/rnaseq-nf"
  input:
    path transcriptome
  output:
    path 'salmon_index'

  script:
  """
  salmon index --threads $task.cpus -t $transcriptome -i salmon_index
  """
}
```

nextflow

Reproducibility

```
process INDEX {
  executor "slurm"
  input:
    path transcriptome
  output:
    path 'salmon_index'

  script:
  """
  salmon index --threads $task.cpus -t $transcriptome -i salmon_index
  """
}
```

nextflow

Reproducible

Between runs

Portable

Between systems

Scalable

Everywhere

nextflow

nextflow

nf-core

nf-core

A community effort to collect a curated set of analysis pipelines built using Nextflow.

https://nf-co.re

nf-core

73

PIPELINES

https://nf-co.re

nf-core

TOOLS

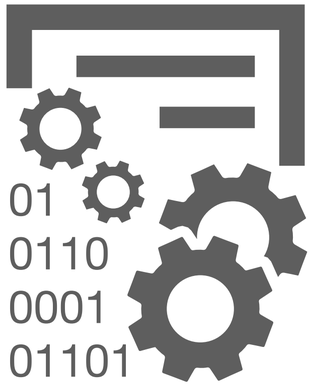
Running pipelines

Writing pipelines

Testing / automation

https://nf-co.re

nf-core



708

MODULES

24

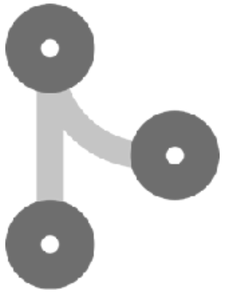
SUB-WORKFLOWS

https://nf-co.re

nf-core

Develop <u>with</u> the community

Use a common template



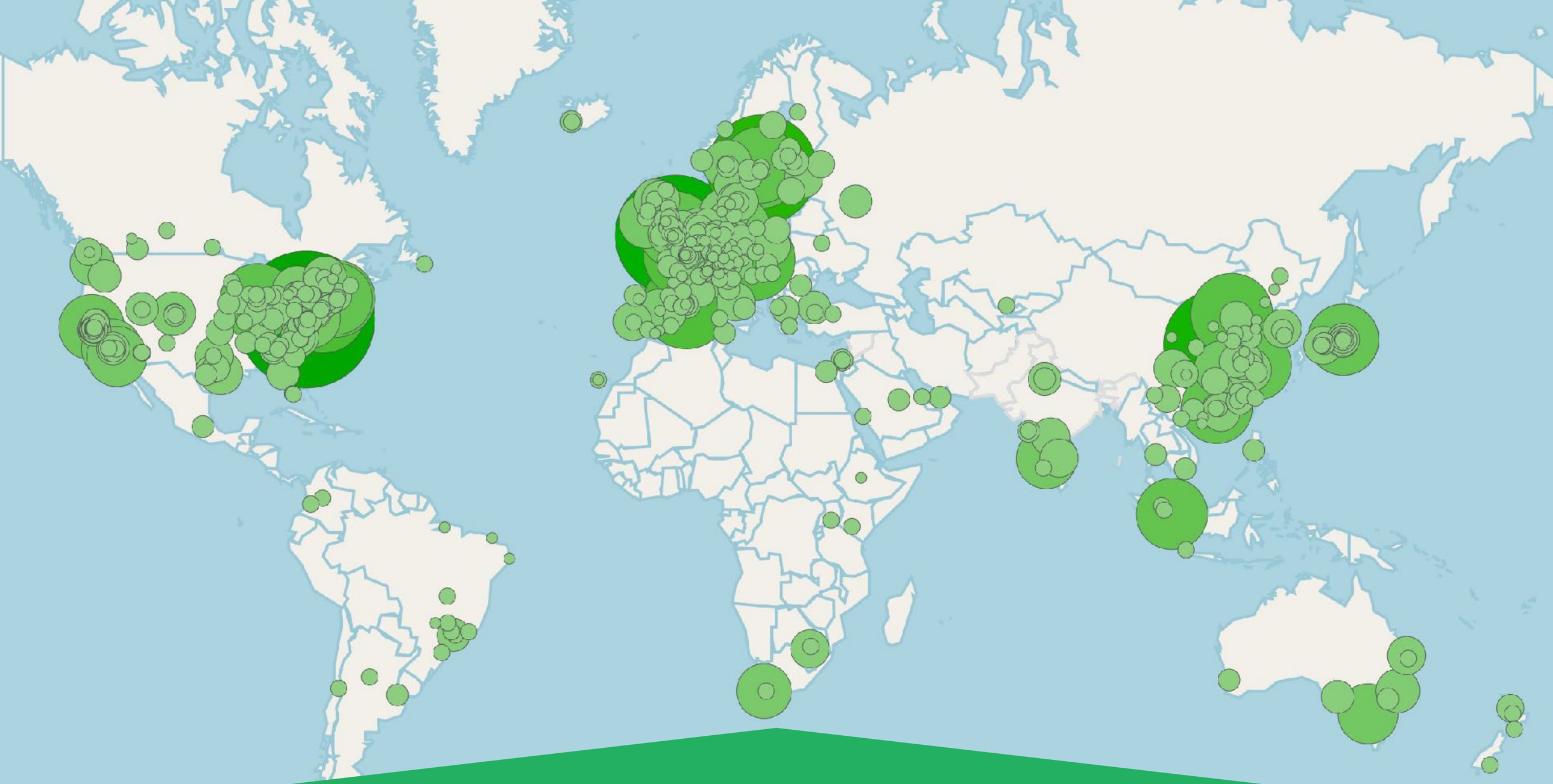
Collaborate, don’t duplicate

https://nf-co.re

4135	527	1583	3294
Slack users	GitHub organisation members	GitHub contributors	Twitter followers

90	11.61K	32.37K	5.21K
Repositories	Pull Requests	Commits	Issues

https://nf-co.re



https://nf-co.re

nature biotechnology

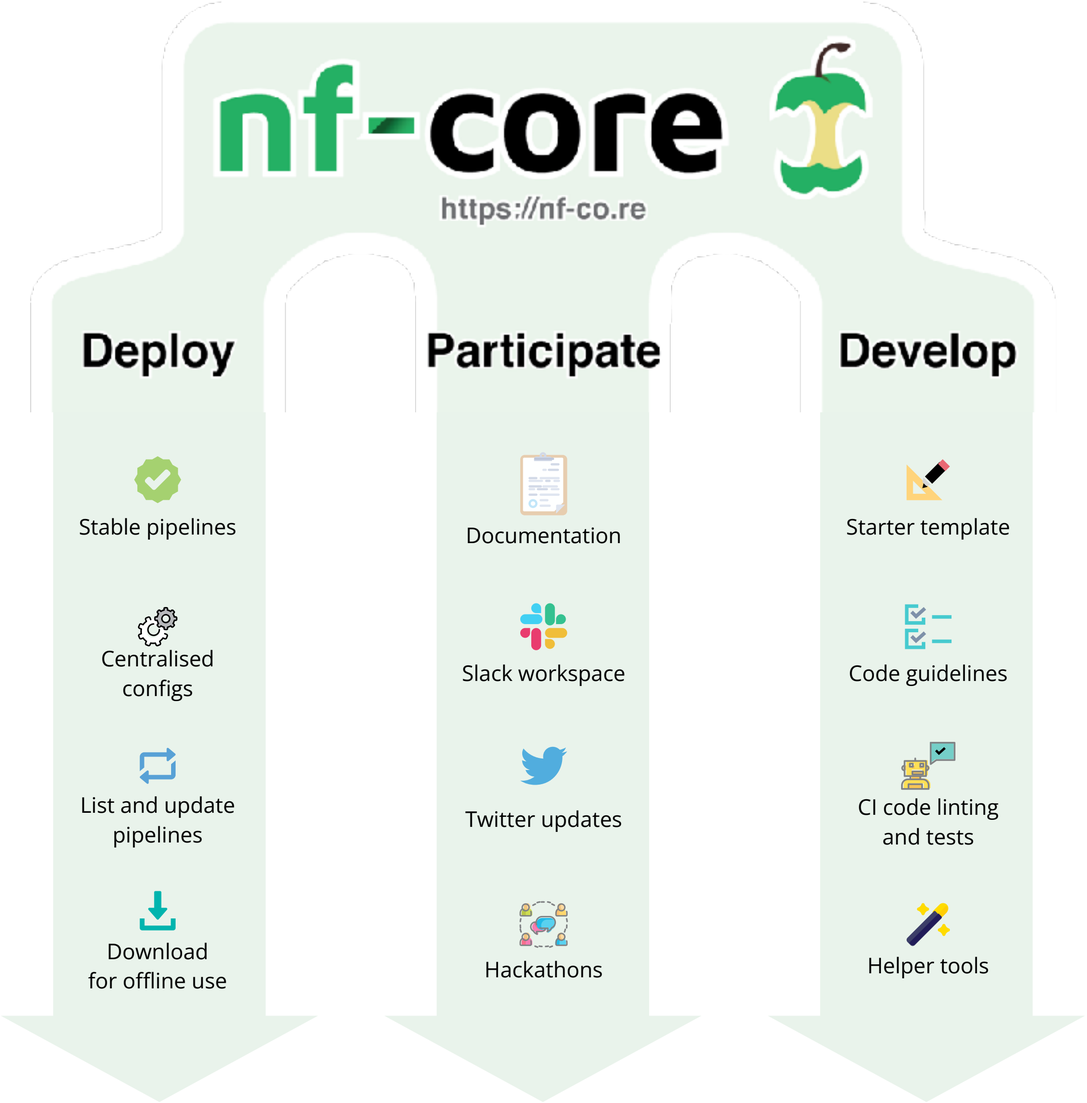
Correspondence | Published: 13 February 2020

The nf-core framework for community-curated bioinformatics pipelines

Philip A. Ewels, Alexander Peltzer, Sven Fillinger, Harshil Patel, Johannes Alneberg, Andreas Wilm, Maxime Ulysse Garcia, Paolo Di Tommaso & Sven Nahnsen

Nature Biotechnology **38**, 276–278(2020) | Cite this article

To the Editor — The standardization, portability and reproducibility of analysis pipelines are key issues within the bioinformatics community. Most bioinformatics pipelines are designed for use on-premises; as a result, the associated software dependencies and execution logic are likely to be tightly coupled with proprietary computing environments. This can make it difficult or even impossible for others to reproduce the ensuing results, which is a fundamental requirement for the validation of scientific findings. Here, we introduce the nf-core framework as a means for the development of collaborative, peer-reviewed, best-practice analysis pipelines (Fig. 1). All nf-core pipelines are written in Nextflow and so inherit the ability to be executed on most computational infrastructures, as well as having native support for container technologies such as Docker and Singularity. The nf-core community (Supplementary Fig. 1) has developed a suite of tools that automate pipeline creation, testing, deployment and synchronization. Our goal is to provide a framework for high-quality bioinformatics pipelines that can be used across all institutions and research facilities.



Join the community

https://nf-co.re/join

nextflow tower

nextflow tower

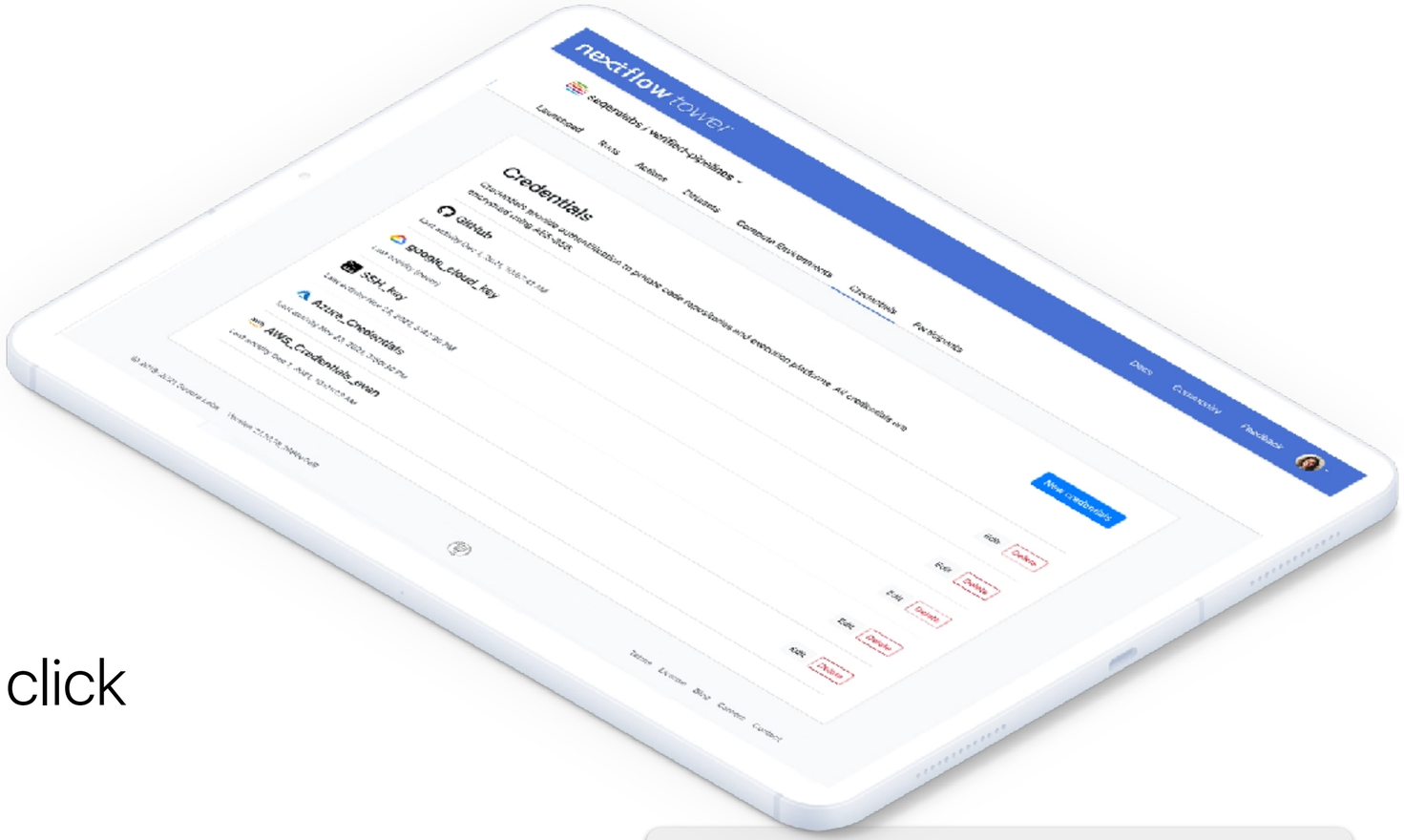
seqeralabs

Intuitive launchpad interface

Launch, manage, and monitor

Share runs and work in teams

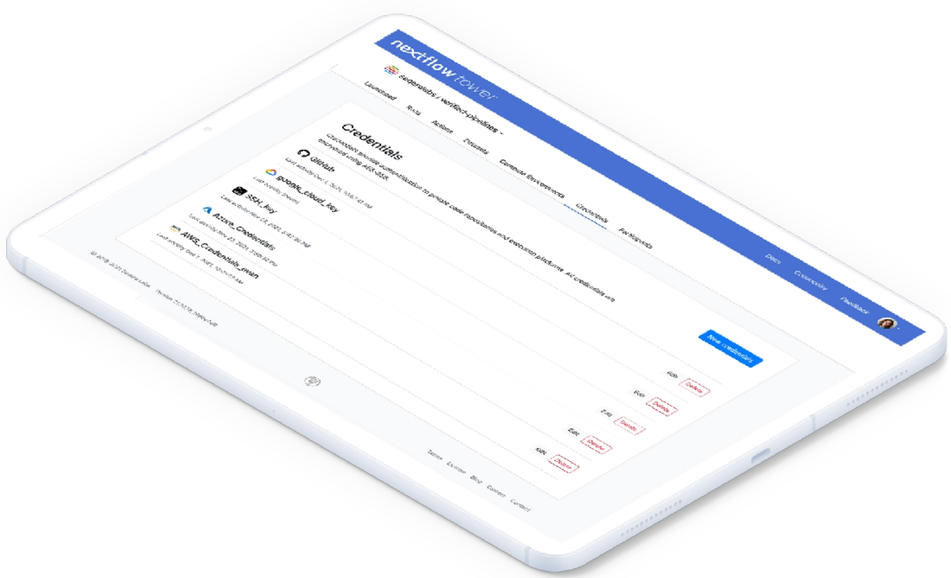
Create cloud infrastructure with a click



https://tower.nf

nextflow tower

seqeralabs



Community: Open source

Cloud: Free & paid tiers

Enterprise: Commercial

`https://tower.nf`

Marcel Ribeiro-Dantas, PhD

seqeralabs
https://seqera.io

http://mribeirodantas.xyz
mribeirodantas
mribeirodantas@seqera.io
mribeirodantas

Chan Zuckerberg Initiative

nextflow
SUMMIT 2022
https://summit.nextflow.io

Nextflow / nf-core training
13-16 March 2023

nf-core hackathon
27-29 March 2023

`https://nf-co.re`

`https://nextflow.io`