nf-core

Reproducible Pipelines for Core Facilities (and you!)

Franziska Bonath ▸ KTH | NGI | SciLifeLab
Maxime U Garcia ▸ Seqera Labs

Outline

- Data Flow to NGI
- QC Steps from Sequencer to Delivery
- Pipelines and Workflow Managers
- nf-core: A Community Curated Set of Pipelines using Nextflow
- Nextflow Pipeline Case Study: Sarek

NGI

(Stockholm node, Illumina projects only)

projects: 631 **samples: 31686**

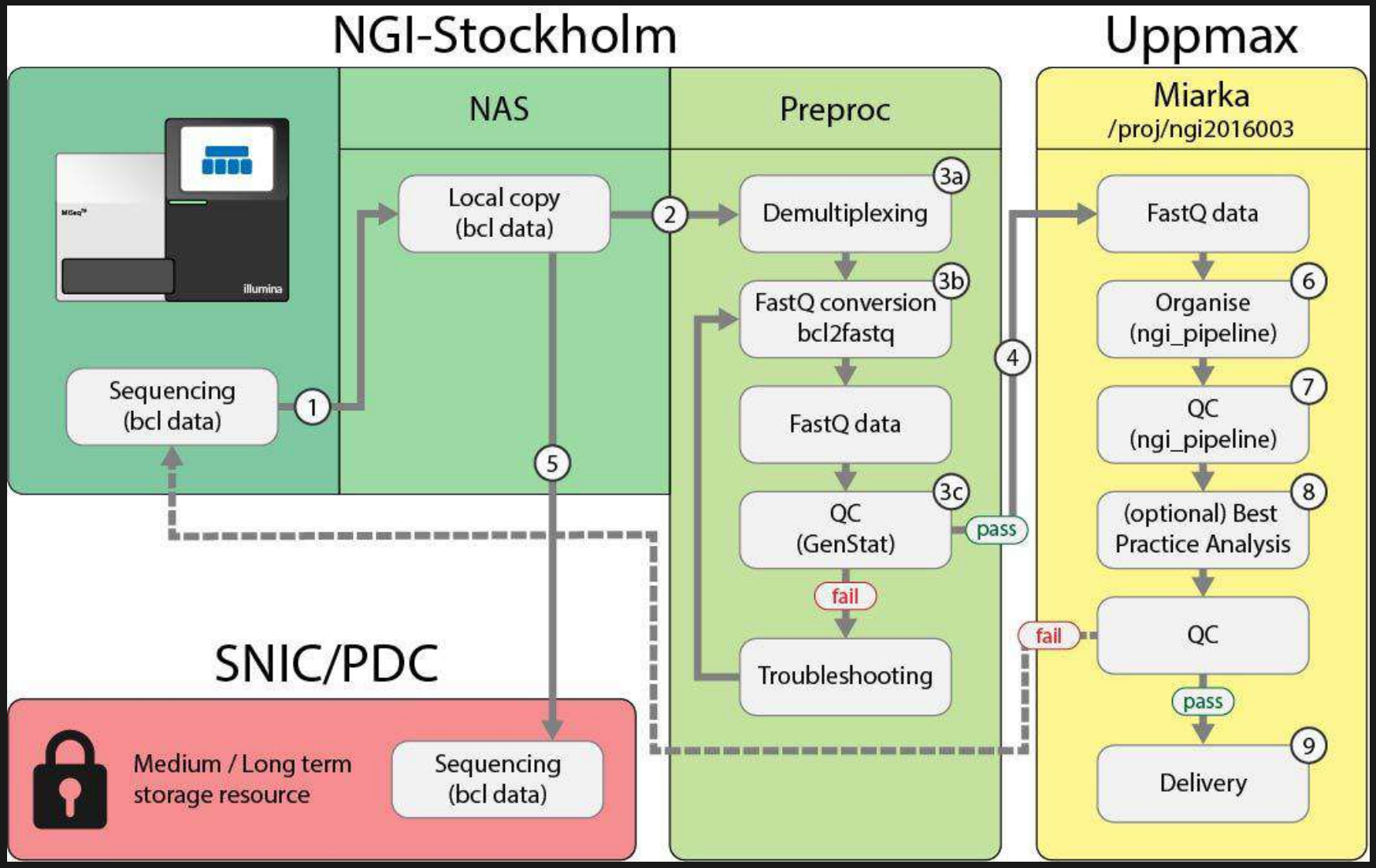
2022

---

bases: 1373 Gbp / day

1 human genome / 3.39 minutes

Data flow at NGI



bioinformatics at NGI

1) Primary QC of flowcells

- Did the flowcell/lane get enough reads?
- Is the average quality of all reads acceptable?
  - % of reads above Phred Q30
  - PhiX error rate below 2 %

Lane 1

Total yield (Mb):	60 687	Total clusters:	518 689 047	% bases > Q30:	95.0	Mean Quality Score:	36.13	% perfect barcode :	99.20	Err. rate	0.1119727622717619

Lane 2

Total yield (Mb):	66 546	Total clusters:	568 769 456	% bases > Q30:	82.54	Mean Quality Score:	33.31	% perfect barcode :	99.48	Err. rate	7.768092488870025

bioinformatics at NGI

2) Demultiplexing

- Did all samples get enough reads
- Are there excessive amounts of undetermined reads?
- Are there valid indexes within the undetermined reads

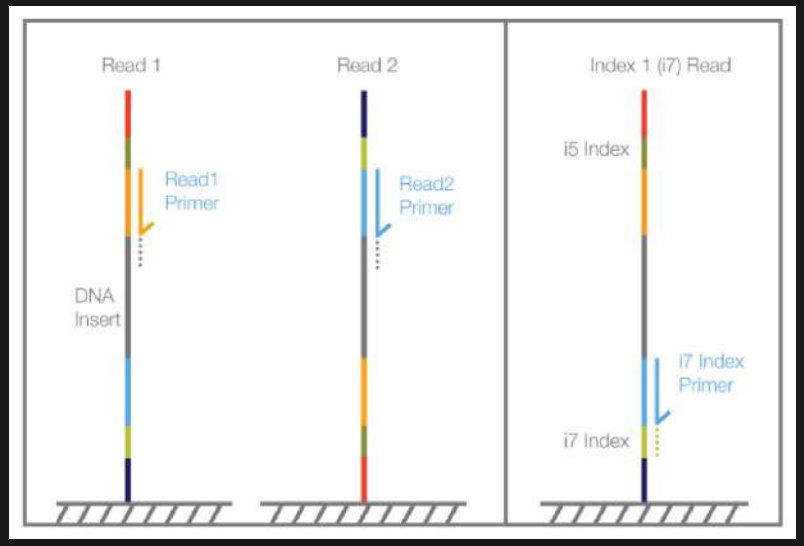
Project Name	Sample Name	Yield (Mb)	Clusters	% BP > Q30	Barcode(s)	Lane %	MQS
E.Xample_23_01	P12345_101	119 875	396 935 437	81.43	AGGCAGAA	96.51	35.45
default	Undetermined	4 334	14 352 127	78.51	unknown	3.49	34.48

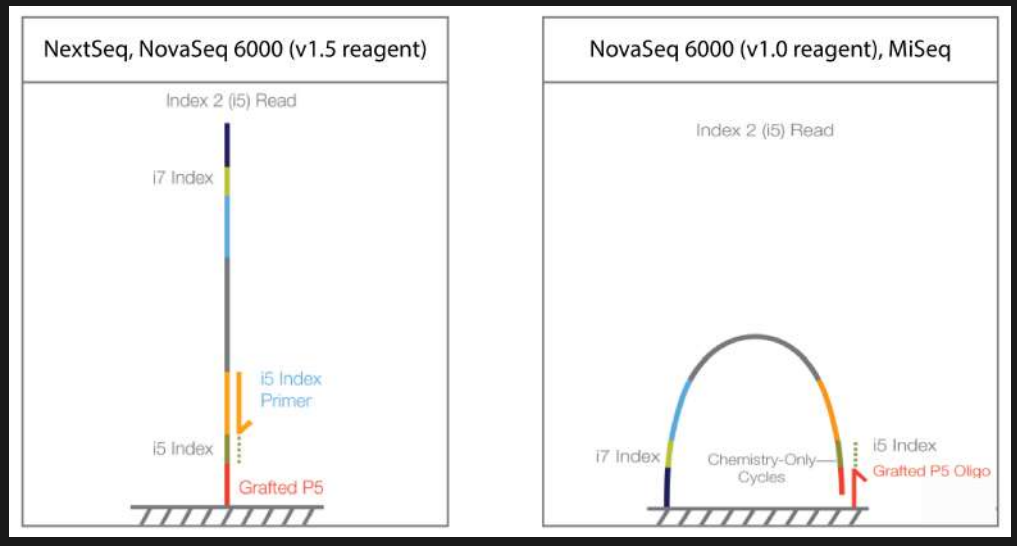
Show Undetermined

Index	Count	Total %	AGGCAGAA mismatches
TGTCTCTT	354 880	2.47%	6
AGCAGAAA	331 060	2.31%	4
CCCCCCCC	278 900	1.94%	7
GGCAGAAA	276 320	1.93%	5

bioinformatics at NGI

2) Demultiplexing



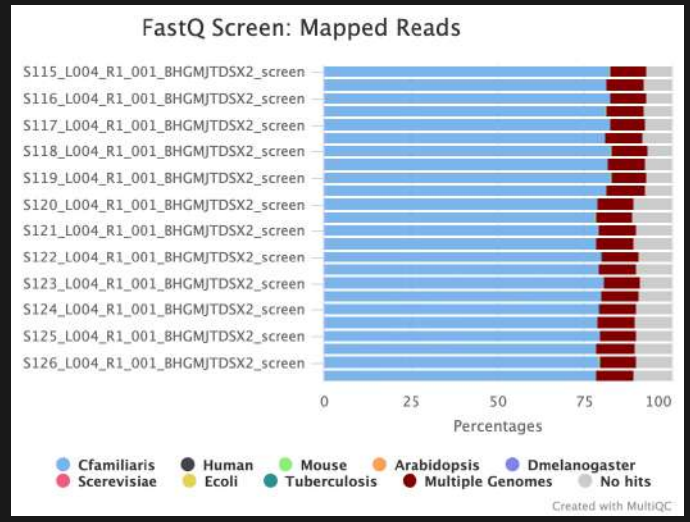


Index Name	i7 Bases in Adapter	i7 Bases for Sample Sheet	i5 Bases in Adapter	i5 Bases for Sample Sheet NovaSeq 6000 with v1.0 reagent kits, MiSeq, HiSeq 2000/2500, NextSeq 2000 (Sample Sheet v2)	i5 Bases for Sample Sheet iSeq, NovaSeq 6000 with v1.5 reagent kits, MiniSeq, NextSeq 500/550, HiSeq 3000/4000/X, NextSeq 2000 (Sample Sheet v1)
UDP0001	CGCTCAGTTC	GAACTGAGCG	TCGTGGAGCG	TCGTGGAGCG	CGCTCCACGA
UDP0002	TATCTGACCT	AGGTCAGATA	CTACAAGATA	CTACAAGATA	TATCTTGTAG
UDP0003	ATATGAGACG	CGTCTCATAT	TATAGTAGCT	TATAGTAGCT	AGCTACTATA

bioinformatics at NGI

3) QC reports by sample

- Quality on sample level
  - % of reads above Phred Q30
- Contamination report (Fastq-screen)
  - mapping against most common species
- Summary of QC report in MultiQC



bioinformatics at NGI

4) "Best Practice" Analysis

- Analysis to control for library preparation issues
- Specific to library preparation type
- First steps of data analysis for the data type
- NGI *cannot* do project specific analysis

- Use of nextflow pipelines under nf-core
- Results are summarized using MultiQC

bioinformatics at NGI

5) Generation of project reports

- Will contain:
  - General QC stats for the flowcell and each sample
  - Information on
    - Library prep
    - Sequencing setup
    - Accreditation status and deviations

bioinformatics at NGI

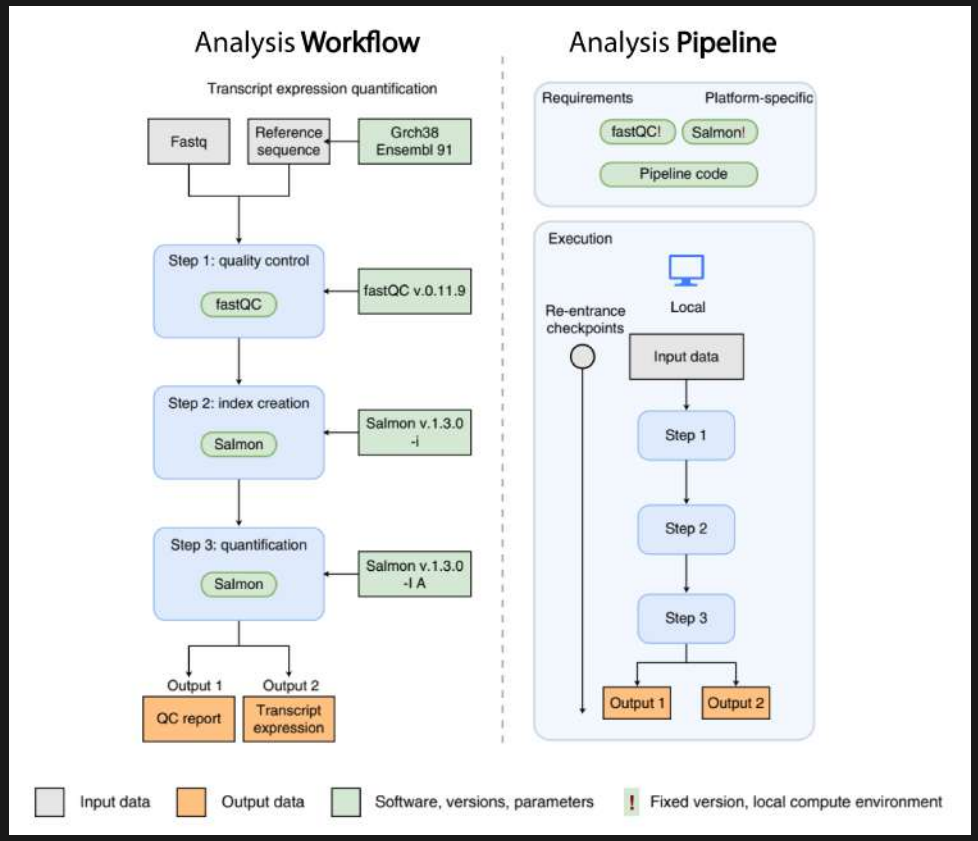
6) Deliveries

- For sensitive data
- Hosted by Uppmax
- Requires a SNIC account

- (Currently) only for non-sensitive data
- hosted by SciLifeLab Data Centre
- Email with access link sent to user

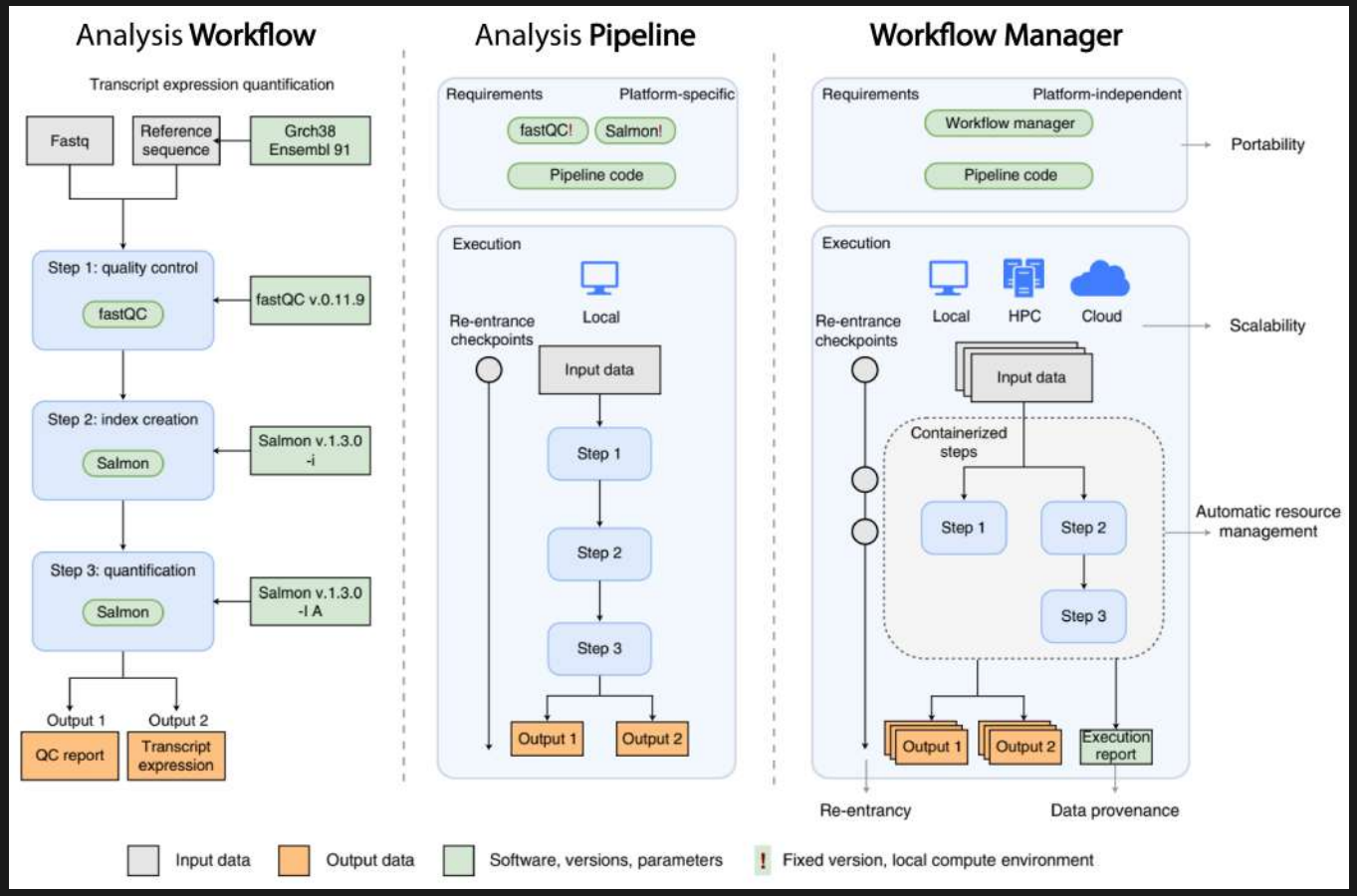
Pipelines

What is a pipeline?



10.1038/s41592-021-01254-9

What is a workflow manager?



10.1038/s41592-021-01254-9

Some available workflow managers

Table 1 | Overview of workflow managers for bioinformatics (top, editable version; bottom, image version)

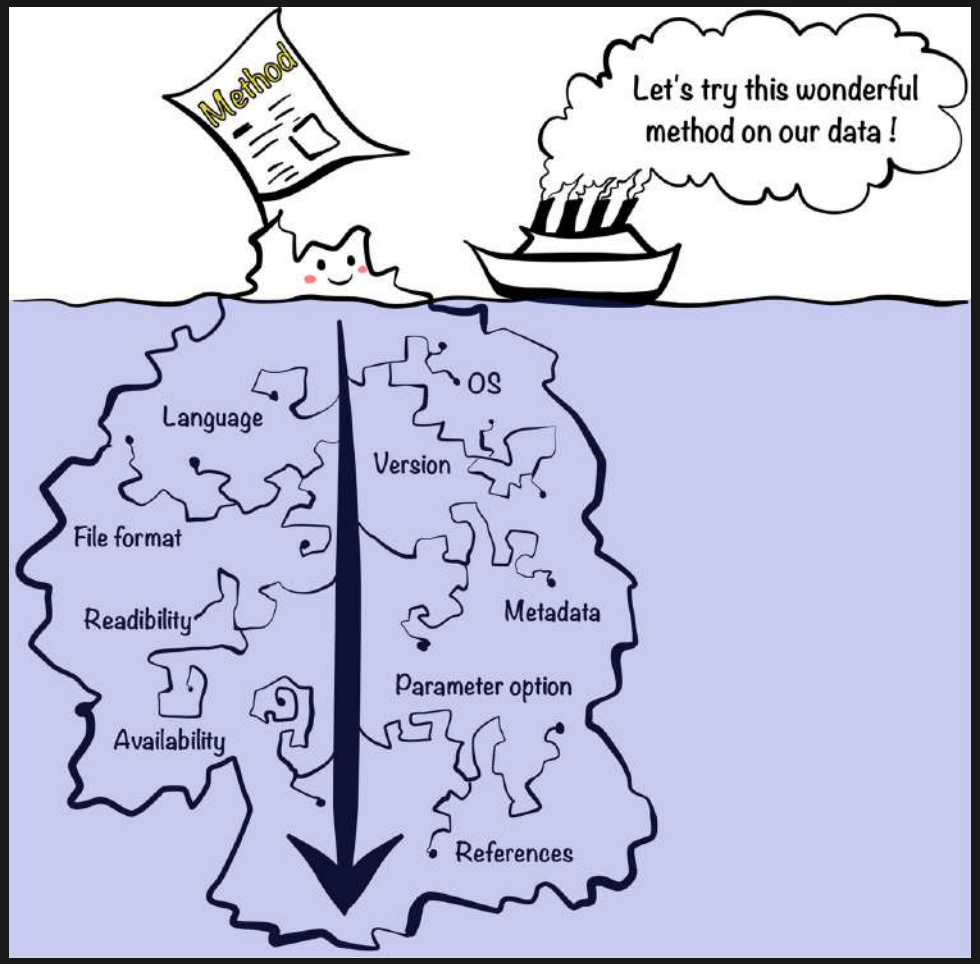
Tool	Class	Ease of use[a]	Expressiveness[b]	Portability[c]	Scalability[d]	Learning resources[e]	Pipeline initiatives[f]
Galaxy	Graphical	●●●	●○○	●●●	●●●	●●●	●●○
KNIME	Graphical	●●●	●○○	○○○	●●◐	●●●	●●○
Nextflow	DSL	●●○	●●●	●●●	●●●	●●●	●●●
Snakemake	DSL	●●○	●●●	●●◐	●●●	●●○	●●●
GenPipes	DSL	●●○	●●●	●●○	●●○	●●○	●●○
bPipe	DSL	●●○	●●●	●●○	●●◐	●●○	●○○
Pachyderm	DSL	●●○	●●●	●○○	●●○	●●●	○○○
SciPipe	Library	●●○	●●●	○○○	○○○	●●○	○○○
Luigi	Library	●●○	●●●	●○○	●●◐	●●○	○○○
Cromwell + WDL	Execution + workflow specification	●○○	●●○	●●●	●●◐	●●○	●●○
cwltool + CWL	Execution + workflow specification	●○○	●●○	●●◐	○○○	●●●	●●○
Toil + CWL/ WDL/Python	Execution + workflow specification	●○○	●●●	●◐○	●●●	●●○	●●○

Please refer to Supplementary Table 1 for details. This information is based on online documentation and manuscripts and may not be reflective of the current state of the projects. Scores for Galaxy are based on the graphical user interface. [a]Ease of use: graphical interface with execution environment (score of 3), programming interface with in-built execution environment (score of 2), separated development and execution environment (score of 1). [b]Expressiveness: based on an existing programming language (3) or a new language or restricted vocabulary (2), primary interaction with graphical user interface (1). [c]Portability: integration with three or more container and package manager platforms (3), two platforms are supported (2), one platform is supported (1). [d]Scalability: considers cloud support, scheduler and orchestration tool integration, and executor support. Please refer to Supplementary Table 1. [e]Learning resources: official tutorials, forums, and events (3), tutorials and forums (2), tutorials or forums (1). [f]Pipelines initiatives: community and curated (3), community or curated (2), not community or curated (1).

10.1038/s41592-021-01254-9

nf-core

Reproducibility is central



10.1093/gigascience/giy077

What is nf-core?

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

What is Nextflow?

- Workflow manager
  - Data driven language
  - Portable
    - executable on multiple platforms
  - Shareable and reproducible
    - with containers or virtual environments

Data driven language

The execution graph depends on the input data,

and is calculated on the go

In `snakemake` it's the other way around

The execution graph depends on the final target,

and is calculated before launch

Portability

www.nextflow.io/docs/latest/executor.html

- Sun Grid Engine, SLURM, PBS/Torque…
- AWS Batch, Kubernetes, Google Life Sciences

Reproducibility

What is nf-core: community

Contributors

The nf-core pipelines and community is driven by many individuals, listed below. This list updates automatically.

Want to see who's working on what? See the contributor leaderboard on the Statistics page.

What is nf-core: for users

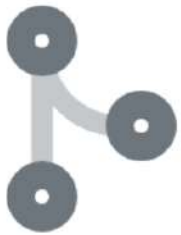
A community effort to collect a curated set of Nextflow pipelines

- Stable releases
- Packaged software
- Documentation
- Portable pipelines
- Continuous integration
- Cloud ready

Ewels, P. A., Peltzer, A., et al. (2020). *Nature Biotechnology*, *38*(3), 276–278.

What is nf-core: for developers

Develop with the community
Join slack, communicate and contribute together to a pipeline



Cooperate, don't duplicate
One pipeline per analysis type, contribute by adding new tools, new features...

What does nf-core provide

- **Pipelines**: ready-made pipelines [n=68]
- **Docs** 🌎: Guidelines, tutorials, videos
- **Subworkflows** 🌎: multi-tool wrappers [n=31]
- **Modules** 🌎: single-tool wrappers [n=797]
- **Configs** 🌎: shared infrastructure configs
- **Test datasets** 🌎: test data for 👆
- **Tools** 🌎: linting, template + automation for 👆

🌎 provided for the larger community

Pipeline requirements

nf-co.re/docs/contributing/adding_pipelines

- Nextflow based
- Common structure
- Stable release tags
- MIT license
- Software bundled for reproducibility
- Continuous Integration testing
- *lagom*

Sarek

nf-co.re/sarek

- Based on GATK Best Practices
- Alignment, Variant Calling, Annotation
- SNPs Indels, SVs, CNV, MSI…
- Germline, Somatic or Tumor only

nf-core/tools

A companion tool

https://nf-co.re/tools

- **launch** - with interactive prompts
- **download** - for offline use
- **lint** - check code against guidelines
- **modules** - List, update, lint, create…
- **subworkflows** - List, update, lint, create…
- …

Configurations

All pipelines come with a default sensible configuration for a regular sized HPC

(Including UPPMAX)

Configurations

github.com/nf-core/configs allows shared configurations between pipelines for a specific HPC

- cpus, time and memory requirements
- scheduler
- queues
- environments
- path to common references files
- …

Training and other events

https://nf-co.re/events

nf-co.re/events/2023/training-march-2023

Need help?

Website: `https://nf-co.re`

Chat: `https://nf-co.re/join`

Follow nf-core on

https://nf-co.re/

Chan Zuckerberg Initiative

Icons: openmoji.org