Introduction to Bioinformatics Workflows with nf-core

Lara Ianov, PhD
Co-Director | UAB Biological Data Science Core (U-BDS)
Assistant Professor | Department of Neurobiology

UAB
Biological
Data
Science
Core

11-19-2025
Cold Spring Harbor Laboratory
Advanced Sequencing Technologies and Bioinformatics

Automating Workflows

- A **workflow** is a series of steps needed to process and/or clean data for easier interpretation
  - Traditionally, computational workflows were written in BASH or Perl
  - As analysis needs grow more complex, workflows require more sophisticated features to complete an analysis

Workflow Management Systems

- **Workflow Management Systems (WfMS)** were developed to meet these complex needs by providing the following:
    - **Environment Management & Portability**
        - WfMS commonly and natively support tools such as *Docker* and *Singularity*
    - **Re-entrancy**
        - WfMS allow users to restart pipelines so that complete steps are "skipped"
    - **Monitoring & Management**
        - WfMS allows users to monitor workflows and provide logs for each step executed
    - **Parallelization & Scalability**
        - WfMS provide methods for steps to be run in parallel

WfMS ensure reproducibility

- **Reproducibility** ensures that analyses always produce the same results, regardless of who is executing the analysis, where its executed, or when its executed

- A common way to ensure reproducibility is through **container technologies** such as *Docker* or *Singularity*

Containers ensure environment reproducibility

- **Containers** are packaged up snapshots of code and/or software and all of its dependencies.
    - Software encapsulated can be executed on any machine afterwards
    - Automate the installation of packages

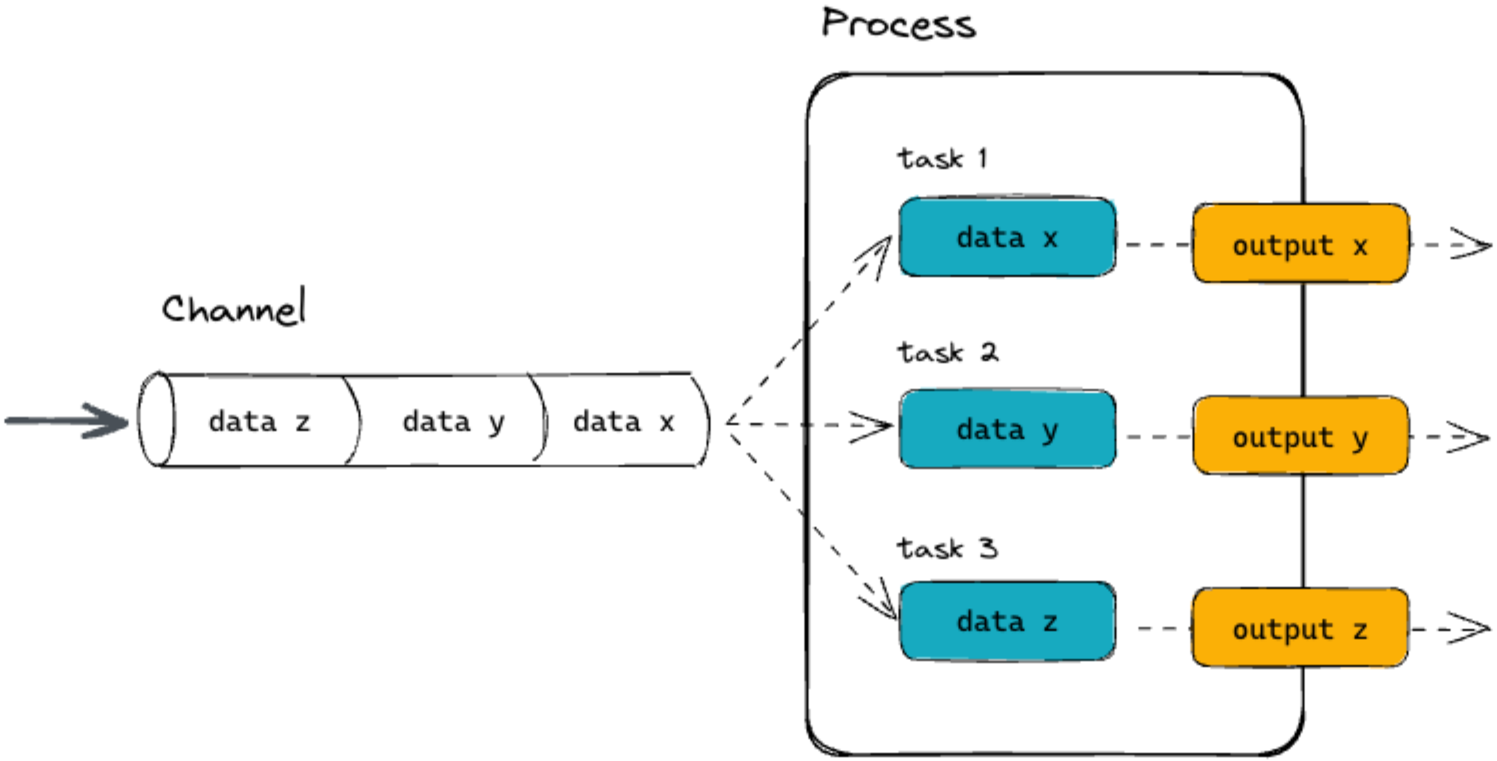
Nextflow is a WfMS

- Nextflow is a WfMS. It uses the **DataFlow Programming Model**
  - **Dataflow Programming** means that each steps waits for the input from a previous step, performs its work, then outputs the result to a downstream step
  - Dataflow Programming ensures Nextflow is high parallelizable, as multiple samples can be in various states during a run

Nextflow is a WfMS

- There are some important aspects and definitions to know
  - Nextflow is written in a language called *Groovy*
  - Nextflow workflows are composed of three parts
    - **Workflow** is the full series of steps to complete an analysis
    - **Channels** contain data produced by processes
    - **Processes** describe the step to be executed, often as a script

Nextflow is a WfMS



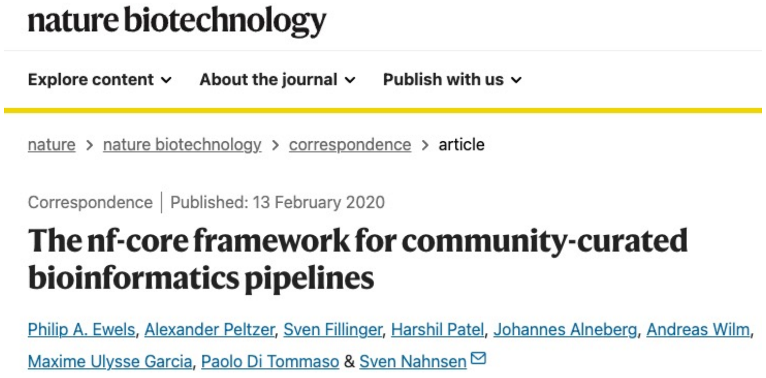
Source: https://training.nextflow.io/2.0/basic_training

Nextflow is a WfMS

- There are some important aspects and definitions to know
    - Nextflow is written in a language called *Groovy*
    - Nextflow workflows are composed of three parts
        - **Workflow** is the full series of steps to complete an analysis
        - **Channels** contain data produced by processes
        - **Processes** describe the step to be executed, often as a script

    - Nextflow can be executed on many platforms, including on the cloud (e.g.: AWS) or High Performance Computing

Nextflow and nf-core



nature biotechnology

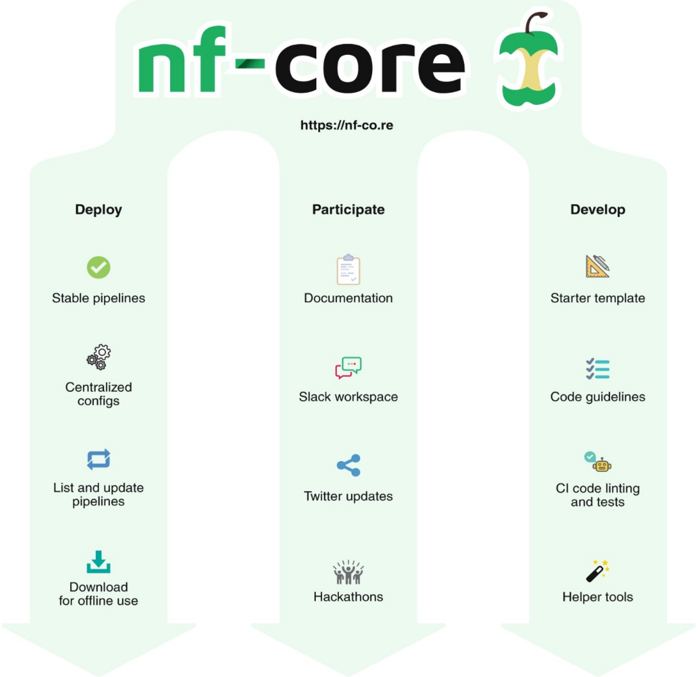
Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature > nature biotechnology > correspondence > article

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



nf-core

https://nf-co.re

Deploy

Stable pipelines

Centralized configs

List and update pipelines

Download for offline use

Participate

Documentation

Slack workspace

Twitter updates

Hackathons

Develop

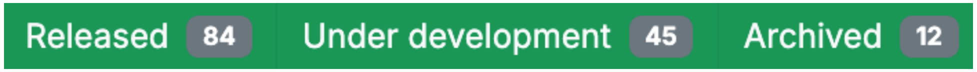
Starter template

Code guidelines

CI code linting and tests

Helper tools

- **nf-core** is a community-led effort to curate, develop, and standardize Nextflow pipelines
- https://nf-co.re/pipelines/
- Current number of pipelines:



Released 84 | Under development 45 | Archived 12

nf-core and reproducibility

nf-core implements several practices to ensure reproducibility

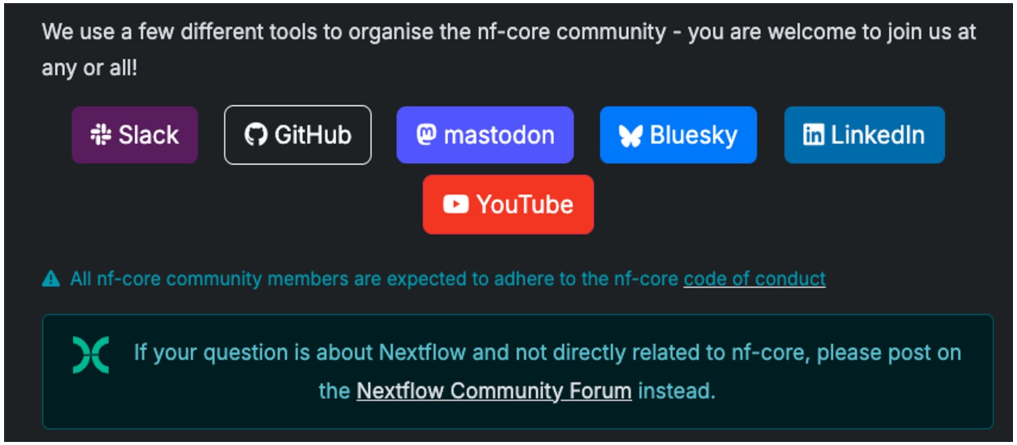
- Nextflow native support (e.g.: containers)
- Version Control - all code is hosted in GitHub
- Pipeline releases at specific points of the pipeline development (initial releases + future support/bug fixes etc.)

nf-core

Nextflow and nf-core

Join nf-core at

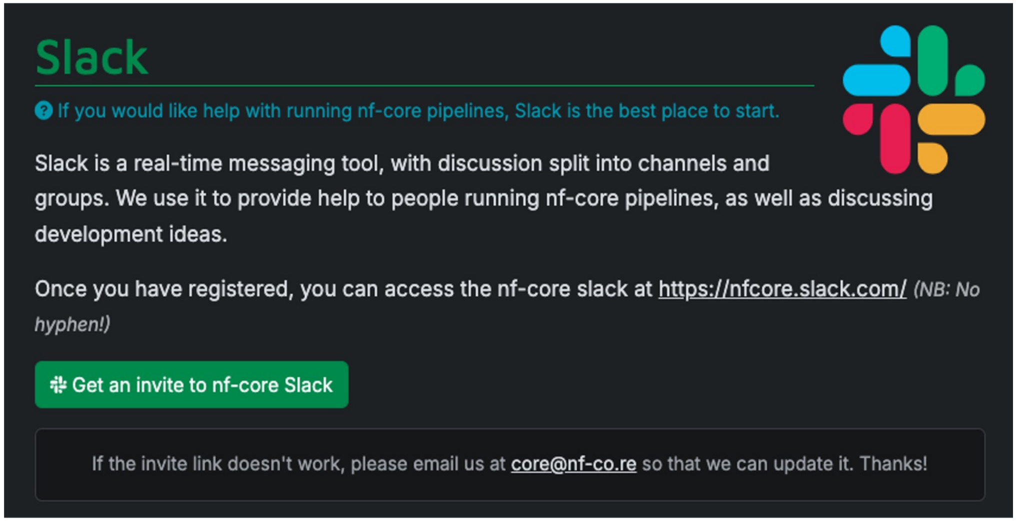
https://nf-co.re/join



We use a few different tools to organise the nf-core community - you are welcome to join us at any or all!

All nf-core community members are expected to adhere to the nf-core code of conduct

If your question is about Nextflow and not directly related to nf-core, please post on the Nextflow Community Forum instead.



Slack

If you would like help with running nf-core pipelines, Slack is the best place to start.

Slack is a real-time messaging tool, with discussion split into channels and groups. We use it to provide help to people running nf-core pipelines, as well as discussing development ideas.

Once you have registered, you can access the nf-core slack at https://nfcore.slack.com/ *(NB: No hyphen!)*

If the invite link doesn't work, please email us at core@nf-co.re so that we can update it. Thanks!

Where to go for help

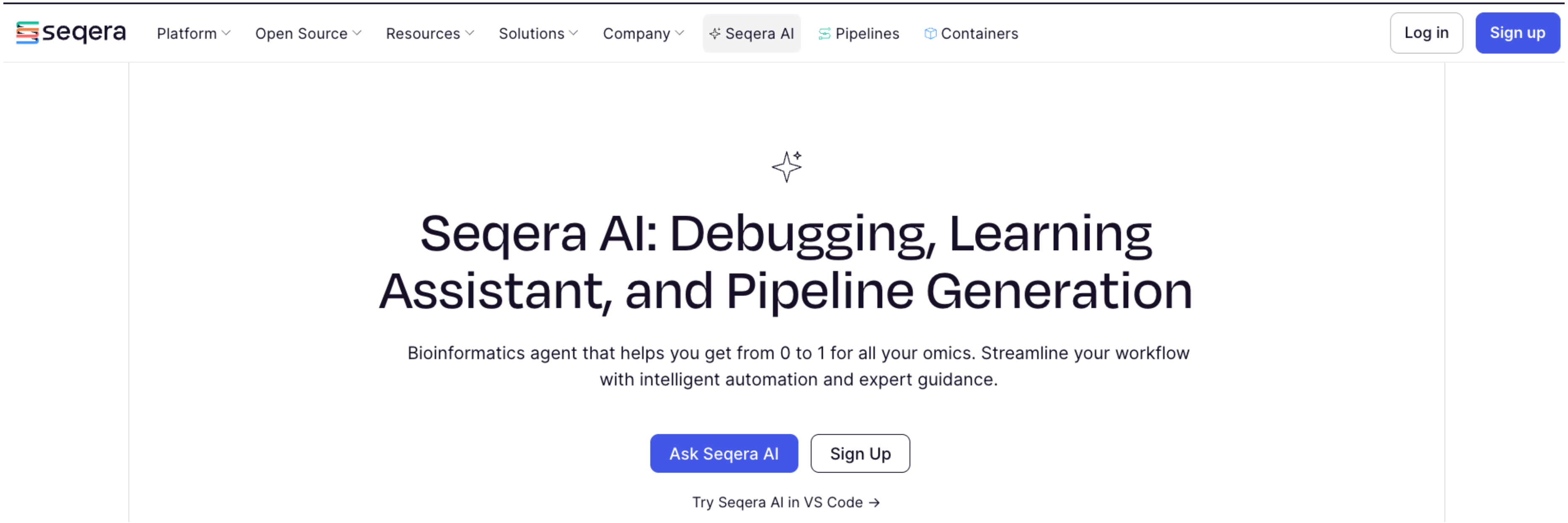
Nextflow

- Slack
- YouTube
- StackOverflow
- SeqeraAI

nf-core

- Slack
- YouTube
- Twitter/X
- Bytesize Talks
- Training Events
- Hackathons
- SeqeraAI

Seqera AI



Seqera AI: Debugging, Learning Assistant, and Pipeline Generation

Bioinformatics agent that helps you get from 0 to 1 for all your omics. Streamline your workflow with intelligent automation and expert guidance.

Seqera AI

Seqera AI online

- Full-featured web environment for pipeline exploration and initial development (including connection to pipelines hosted in GitHub)
- Ability to test code snippets and validate pipeline components
- SRA dataset search with natural language queries

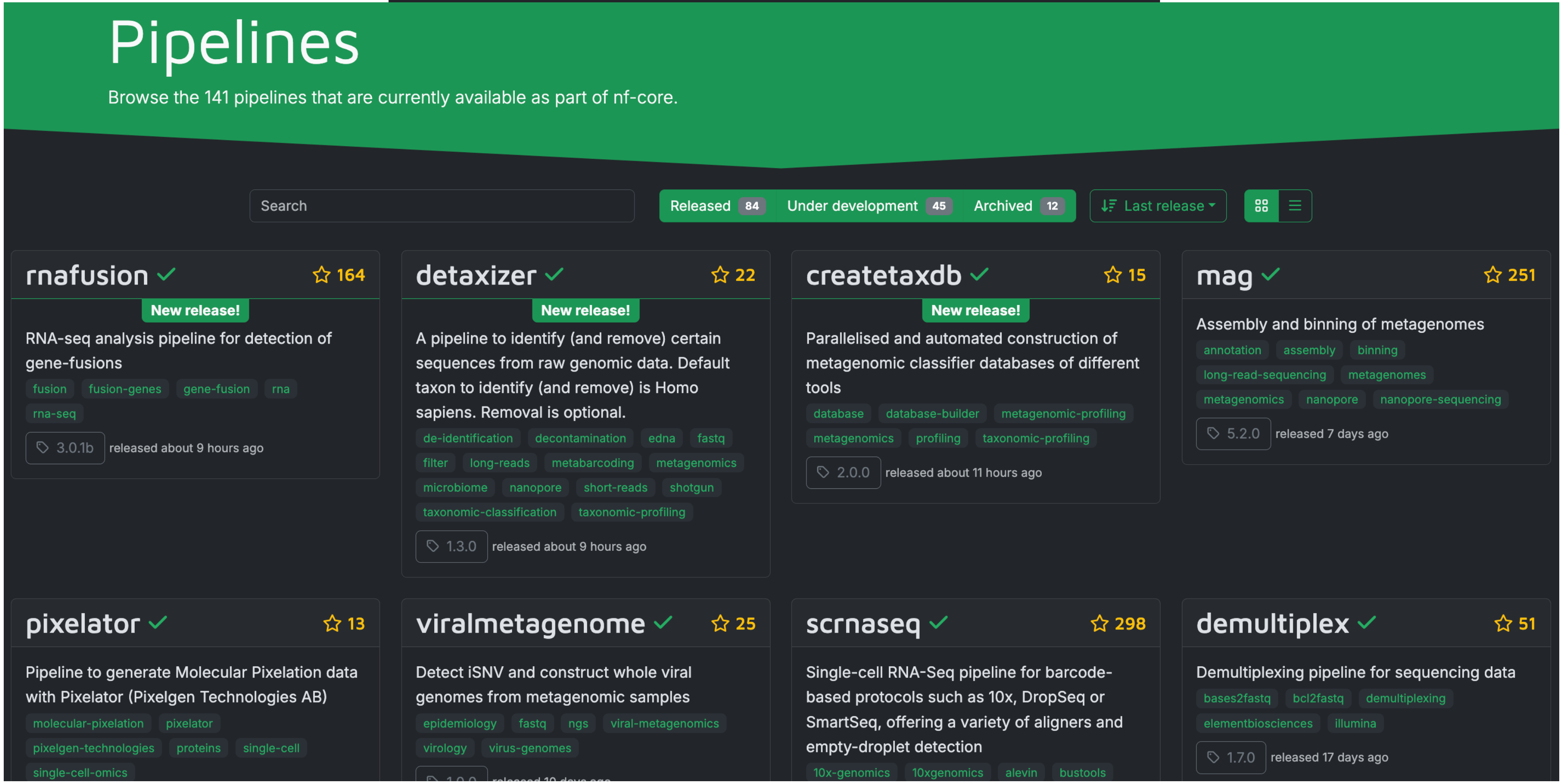
Nextflow VS Code Extension

- Provides IDE-native experience which can facilitate in-depth pipeline development
- Real-time support for coding
- Direct access to log files and terminal outputs can enhance debugging

Users will get the most benefit when combining Seqera AI with an understanding of the underlying pipeline / by having foundational knowledge.

nf-core

A global community collaborating to build open-source Nextflow components and pipelines



Pipelines

Browse the 141 pipelines that are currently available as part of nf-core.

rnafusion ✓ ☆ 164

New release!

RNA-seq analysis pipeline for detection of gene-fusions

fusion · fusion-genes · gene-fusion · rna · rna-seq

3.0.1b released about 9 hours ago

detaxizer ✓ ☆ 22

New release!

A pipeline to identify (and remove) certain sequences from raw genomic data. Default taxon to identify (and remove) is Homo sapiens. Removal is optional.

de-identification · decontamination · edna · fastq · filter · long-reads · metabarcoding · metagenomics · microbiome · nanopore · short-reads · shotgun · taxonomic-classification · taxonomic-profiling

1.3.0 released about 9 hours ago

createtaxdb ✓ ☆ 15

New release!

Parallelised and automated construction of metagenomic classifier databases of different tools

database · database-builder · metagenomic-profiling · metagenomics · profiling · taxonomic-profiling

2.0.0 released about 11 hours ago

mag ✓ ☆ 251

Assembly and binning of metagenomes

annotation · assembly · binning · long-read-sequencing · metagenomes · metagenomics · nanopore · nanopore-sequencing

5.2.0 released 7 days ago

pixelator ✓ ☆ 13

Pipeline to generate Molecular Pixelation data with Pixelator (Pixelgen Technologies AB)

molecular-pixelation · pixelator · pixelgen-technologies · proteins · single-cell · single-cell-omics

viralmetagenome ✓ ☆ 25

Detect iSNV and construct whole viral genomes from metagenomic samples

epidemiology · fastq · ngs · viral-metagenomics · virology · virus-genomes

1.0.0 released 10 days ago

scrnaseq ✓ ☆ 298

Single-cell RNA-Seq pipeline for barcode-based protocols such as 10x, DropSeq or SmartSeq, offering a variety of aligners and empty-droplet detection

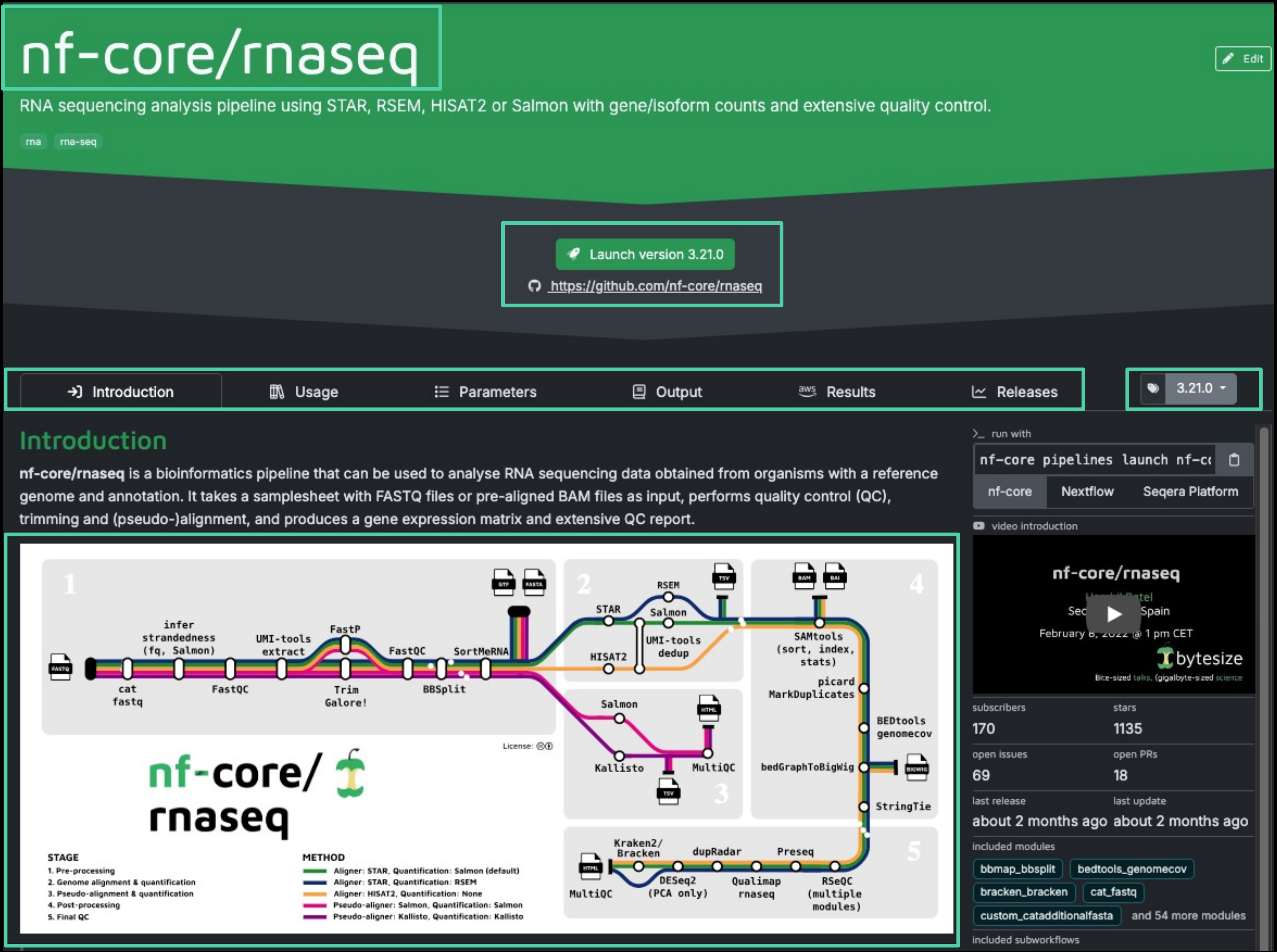
10x-genomics · 10xgenomics · alevin · bustools

demultiplex ✓ ☆ 51

Demultiplexing pipeline for sequencing data

bases2fastq · bcl2fastq · demultiplexing · elementbiosciences · illumina

1.7.0 released 17 days ago



nf-core/rnaseq

RNA sequencing analysis pipeline using STAR, RSEM, HISAT2 or Salmon with gene/isoform counts and extensive quality control.

rna rna-seq

Introduction

nf-core/rnaseq is a bioinformatics pipeline that can be used to analyse RNA sequencing data obtained from organisms with a reference genome and annotation. It takes a samplesheet with FASTQ files or pre-aligned BAM files as input, performs quality control (QC), trimming and (pseudo-)alignment, and produces a gene expression matrix and extensive QC report.

STAGE
1. Pre-processing
2. Genome alignment & quantification
3. Pseudo-alignment & quantification
4. Post-processing
5. Final QC

METHOD
- Aligner: STAR, Quantification: Salmon (default)
- Aligner: STAR, Quantification: RSEM
- Aligner: HISAT2, Quantification: None
- Pseudo-aligner: Salmon, Quantification: Salmon
- Pseudo-aligner: Kallisto, Quantification: Kallisto

run with

```
nf-core pipelines launch nf-cc
```

video introduction

nf-core/rnaseq
Harshil Patel
Sec Spain
February 8, 2022 @ 1 pm CET
bytesize
Bite-sized talks. (gigalbyte-sized science

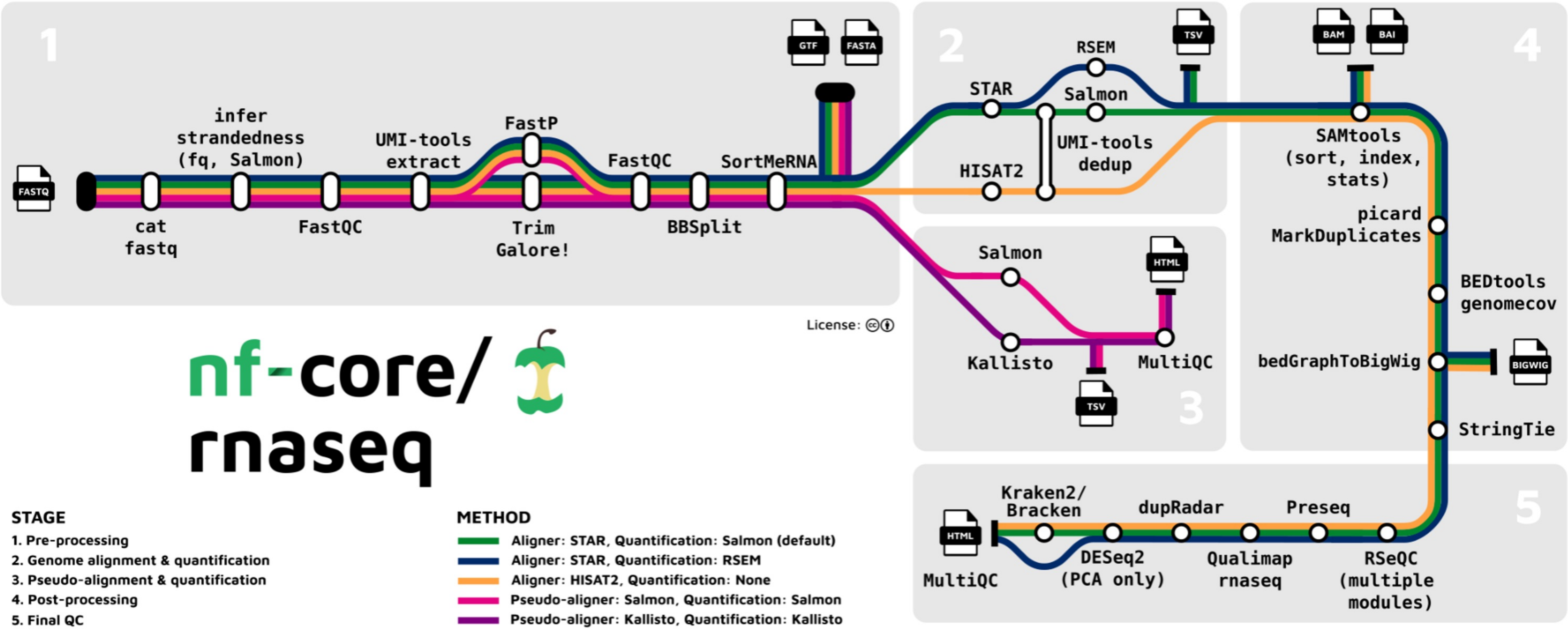
subscribers	stars
170	1135
open issues	open PRs
69	18
last release	last update
about 2 months ago	about 2 months ago

included modules

bbmap_bbsplit bedtools_genomecov bracken_bracken cat_fastq custom_catadditionalfasta and 54 more modules

included subworkflows

RNA-Seq Pipeline



STAGE

1. Pre-processing
2. Genome alignment & quantification
3. Pseudo-alignment & quantification
4. Post-processing
5. Final QC

METHOD

- 🟩 Aligner: STAR, Quantification: Salmon (default)
- 🟦 Aligner: STAR, Quantification: RSEM
- 🟧 Aligner: HISAT2, Quantification: None
- 🟥 Pseudo-aligner: Salmon, Quantification: Salmon
- 🟪 Pseudo-aligner: Kallisto, Quantification: Kallisto

https://nf-co.re/rnaseq/

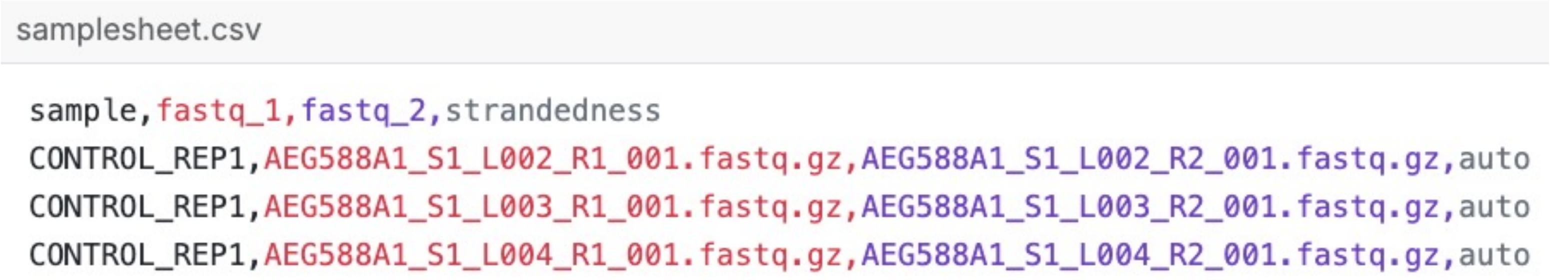
RNA-Seq Pipeline

- nf-core/rnaseq is configured only for **short read** sequencing data
- Long read sequencing data requires additional analytical steps, and not all tools made for short read data will work for long read data
  - Many of the concepts discussed today can be applied to more dedicated pipelines such as **nf-core/nanoseq**

Anatomy of nf-core workflows implementation

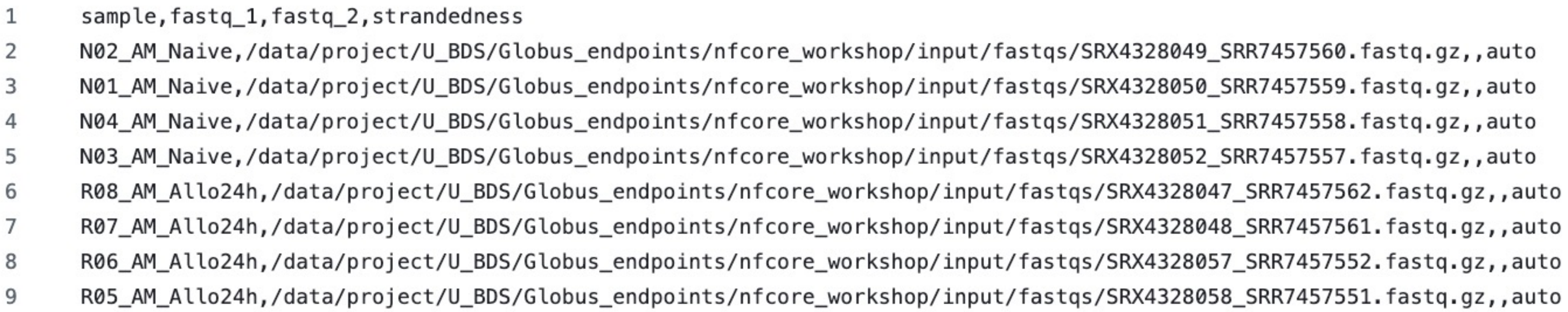
- All contain a small test dataset, the “test profile”
  - **-profile** test
- Submitting workflows, include:
  - **samplesheet.csv**
  - **Selection of pipeline specific parameters**
    - **CLI** or Nextflow **-params-file** option
  - **Submission script to run the Nextflow runner job**
    - Overall structure similar but will differ depending on your choice where to run
      - E.g.: Docker vs Singularity

RNA-Seq Pipeline – samplesheet.csv



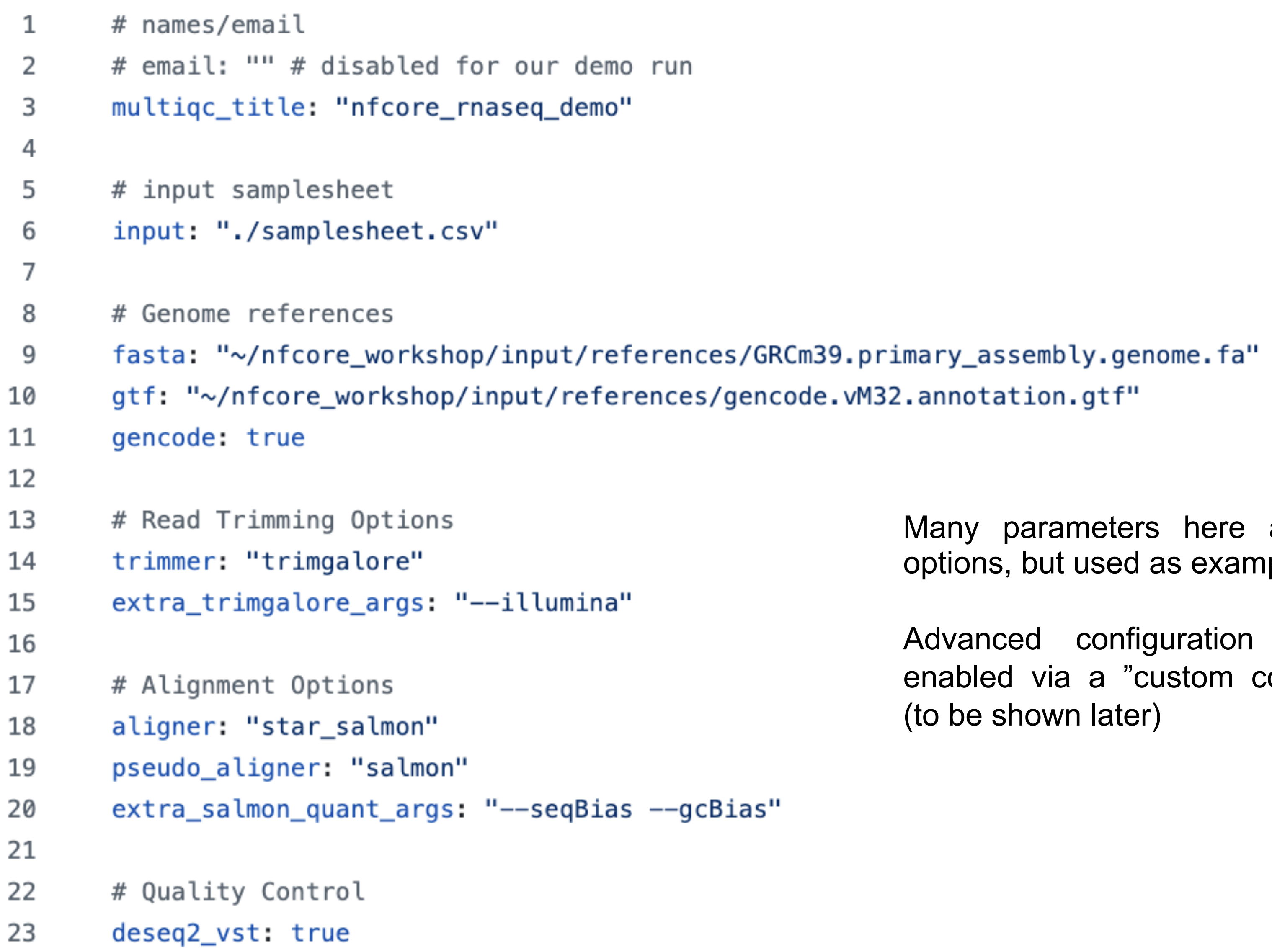
samplesheet.csv

```
sample,fastq_1,fastq_2,strandedness
CONTROL_REP1,AEG588A1_S1_L002_R1_001.fastq.gz,AEG588A1_S1_L002_R2_001.fastq.gz,auto
CONTROL_REP1,AEG588A1_S1_L003_R1_001.fastq.gz,AEG588A1_S1_L003_R2_001.fastq.gz,auto
CONTROL_REP1,AEG588A1_S1_L004_R1_001.fastq.gz,AEG588A1_S1_L004_R2_001.fastq.gz,auto
```



```
sample,fastq_1,fastq_2,strandedness
N02_AM_Naive,/data/project/U_BDS/Globus_endpoints/nfcore_workshop/input/fastqs/SRX4328049_SRR7457560.fastq.gz,,auto
N01_AM_Naive,/data/project/U_BDS/Globus_endpoints/nfcore_workshop/input/fastqs/SRX4328050_SRR7457559.fastq.gz,,auto
N04_AM_Naive,/data/project/U_BDS/Globus_endpoints/nfcore_workshop/input/fastqs/SRX4328051_SRR7457558.fastq.gz,,auto
N03_AM_Naive,/data/project/U_BDS/Globus_endpoints/nfcore_workshop/input/fastqs/SRX4328052_SRR7457557.fastq.gz,,auto
R08_AM_Allo24h,/data/project/U_BDS/Globus_endpoints/nfcore_workshop/input/fastqs/SRX4328047_SRR7457562.fastq.gz,,auto
R07_AM_Allo24h,/data/project/U_BDS/Globus_endpoints/nfcore_workshop/input/fastqs/SRX4328048_SRR7457561.fastq.gz,,auto
R06_AM_Allo24h,/data/project/U_BDS/Globus_endpoints/nfcore_workshop/input/fastqs/SRX4328057_SRR7457552.fastq.gz,,auto
R05_AM_Allo24h,/data/project/U_BDS/Globus_endpoints/nfcore_workshop/input/fastqs/SRX4328058_SRR7457551.fastq.gz,,auto
```

RNA-Seq Pipeline – params.yml

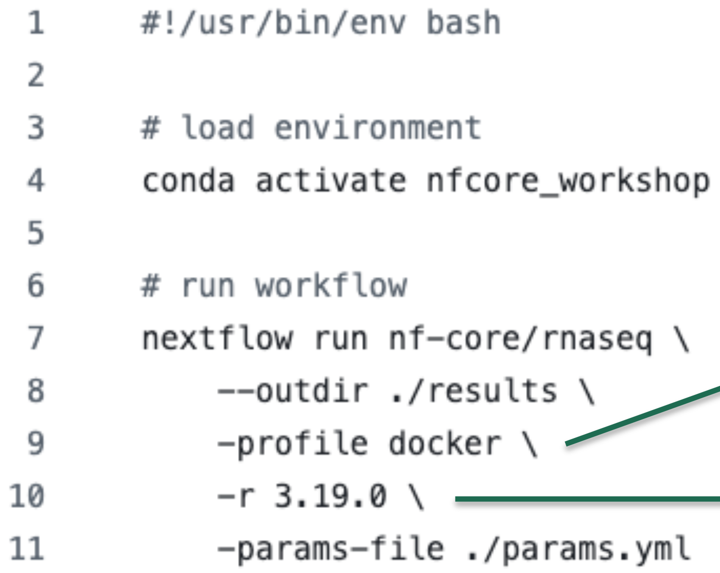


```
# names/email
# email: "" # disabled for our demo run
multiqc_title: "nfcore_rnaseq_demo"

# input samplesheet
input: "./samplesheet.csv"

# Genome references
fasta: "~/nfcore_workshop/input/references/GRCm39.primary_assembly.genome.fa"
gtf: "~/nfcore_workshop/input/references/gencode.vM32.annotation.gtf"
gencode: true

# Read Trimming Options
trimmer: "trimgalore"
extra_trimgalore_args: "--illumina"

# Alignment Options
aligner: "star_salmon"
pseudo_aligner: "salmon"
extra_salmon_quant_args: "--seqBias --gcBias"

# Quality Control
deseq2_vst: true
```

Many parameters here are the default options, but used as examples.

Advanced configuration can also be enabled via a ”custom configuration” file (to be shown later)

RNA-Seq Pipeline – submission script



```
#!/usr/bin/env bash

# load environment
conda activate nfcore_workshop

# run workflow
nextflow run nf-core/rnaseq \
    --outdir ./results \
    -profile docker \
    -r 3.19.0 \
    -params-file ./params.yml
```

Conda env: **Nextflow** is the key dependency; **nf-core/tools** required for developers

Docker profile for local computer (or any other sudo privilege environment)

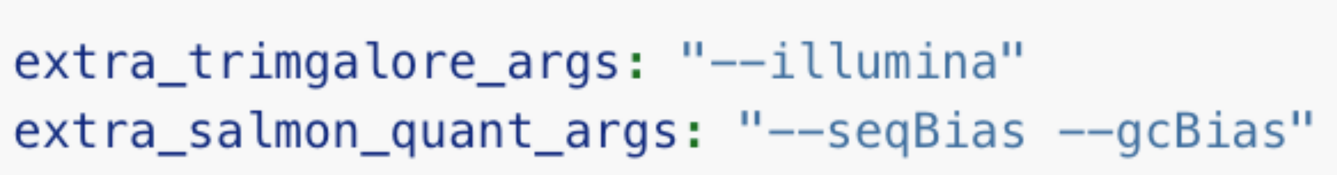
Specifying the version is highly recommend (even when using the latest)

For HPC enable singularity profile OR an institutional profile if your institution has one (you can create one yourself as well)

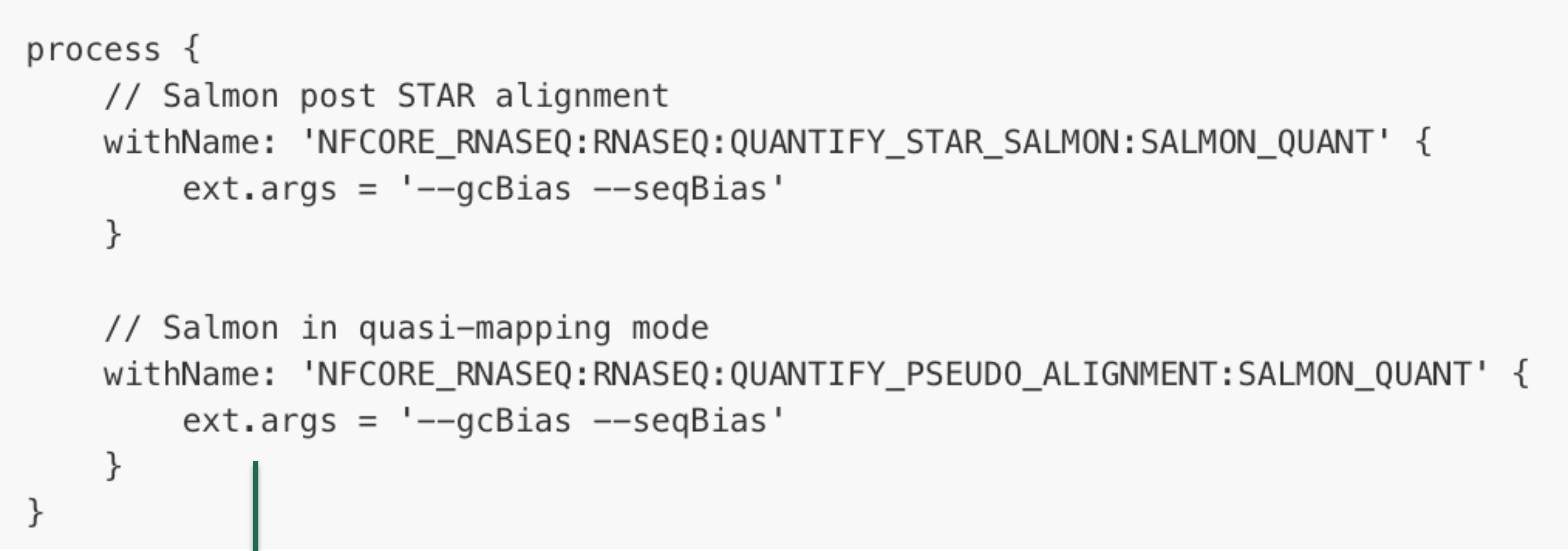
Configuration files

- Can be used to modify tool-specific parameters for **any Nextflow pipeline** or other workflow configuration – e.g.: computational resources
- Passed to the pipeline via the **–c <file_name>** (e.g.: file_name = custom.config)

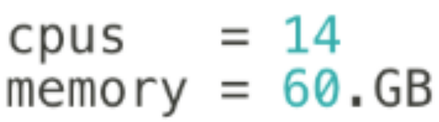
Configuration files



```
extra_trimgalore_args: "--illumina"
extra_salmon_quant_args: "--seqBias --gcBias"
```



```
process {
    // Salmon post STAR alignment
    withName: 'NFCORE_RNASEQ:RNASEQ:QUANTIFY_STAR_SALMON:SALMON_QUANT' {
        ext.args = '--gcBias --seqBias'
    }

    // Salmon in quasi-mapping mode
    withName: 'NFCORE_RNASEQ:RNASEQ:QUANTIFY_PSEUDO_ALIGNMENT:SALMON_QUANT' {
        ext.args = '--gcBias --seqBias'
    }
}
```



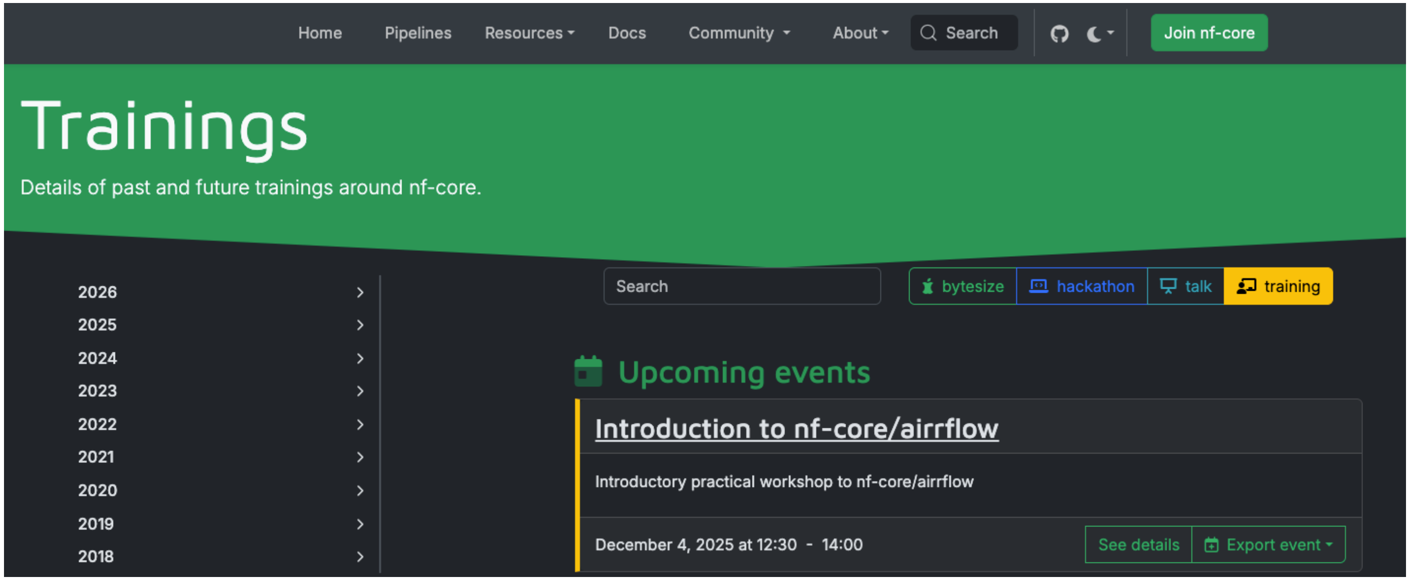
```
cpus   = 14
memory = 60.GB
```

Computational resources can be added in the same manner (**cpus, memory, time**)

Salmon doesn't need this much, just an example ☺

Learning more

- Slack, YouTube channels, Training Week, SeqeraAI etc.
- Detailed tutorial from scripts shown here:
- https://u-bds.github.io/nf-core_workshop
- nf-core training website (**Community** → **Training**)



Trainings

Details of past and future trainings around nf-core.

Upcoming events

Introduction to nf-core/airrflow

Introductory practical workshop to nf-core/airrflow

December 4, 2025 at 12:30 - 14:00

Questions

UAB Biological Data Science Core

Website: https://www.uab.edu/cores/ircp/bds | **X:** @UAB_BDS