Tufts UNIVERSITY

Reproducible, scalable bioinformatics workflows with nextflow and nf-core

Shirley (Xue) Li
Bioinformatician
TTS Research Technology

Yucheng Zhang
Bioinformatics Engineer
TTS Research Technology

tts-research@tufts.edu

nextflow

nf-core

OPEN OnDemand

Overview

- Day 1 (April 3)
  1. Intro to nextflow and nf-core (Yucheng)
  2. How to run nf-core pipelines at Tufts HPC (Yucheng)
  3. How to download raw fastq data with nf-core fetchngs pipeline (Shirley)

- Day 2 (April 4)
  1. Clean cache data (Yucheng)
  2. Nextflow tower (Yucheng)
  3. Running RNA-Seq analysis with nf-core rnaseq pipeline (Shirley)

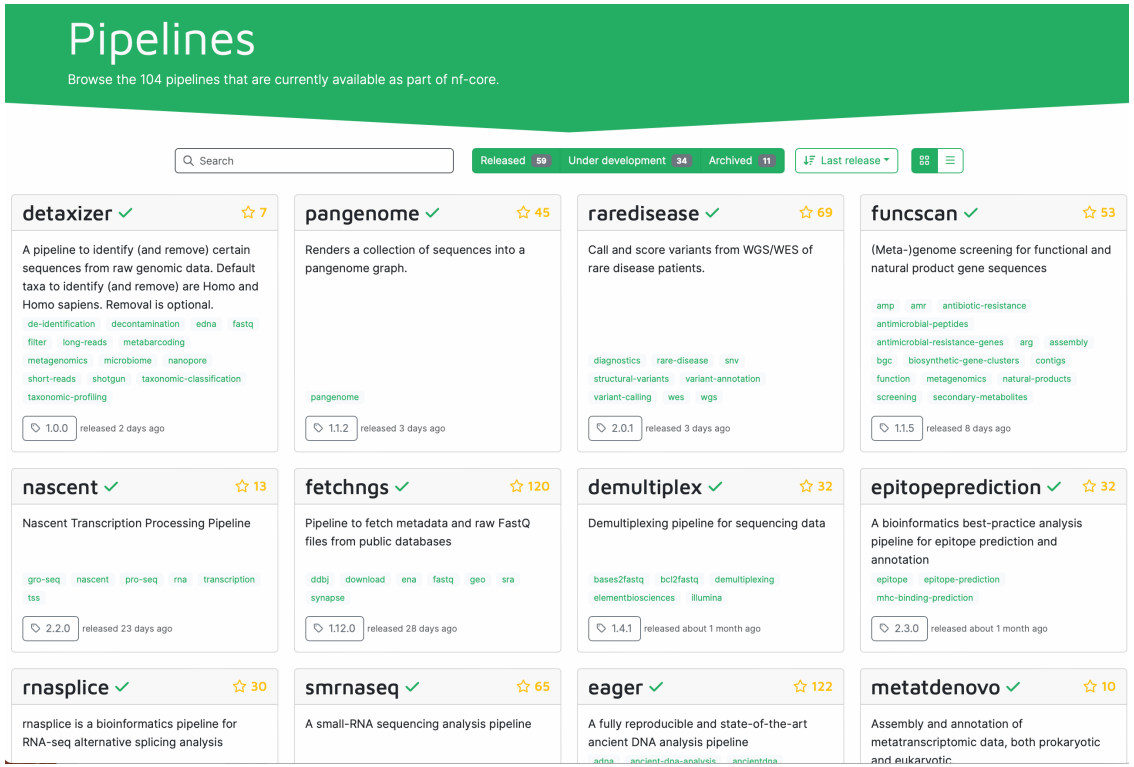
- Day 3 (April 11)
  1. RNA-seq downstream analysis with nf-core differentialabundance pipeline (Shirley)
  2. Visualize outputs with shinyNGS (Shirley)
  3. Troubleshooting (Yucheng)

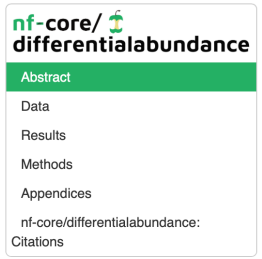
Office Hour

Apr 5 & 12, 1-3pm

- **Tisch Library, Room 208A**

What to expect from this workshop





nf-core/differentialabundance

Differential gene abundance report: PRMT5kd vs. GFPkd

By Yucheng Zhang, differentialabundance workflow version1.4.0

2024-03-27

Abstract

This report summarises differential gene analysis as performed by the nf-core/differentialabundance pipeline.

Data

Samples

A summary of sample metadata is below:

Sample metadata

Sample	Treatment	Replicate	Batch
GFPkd_1	GFPkd	1	A
GFPkd_2	GFPkd	2	A
GFPkd_3	GFPkd	3	A
PRMT5kd_1	PRMT5kd	1	A
PRMT5kd_2	PRMT5kd	2	A
PRMT5kd_3	PRMT5kd	3	A

Contrasts

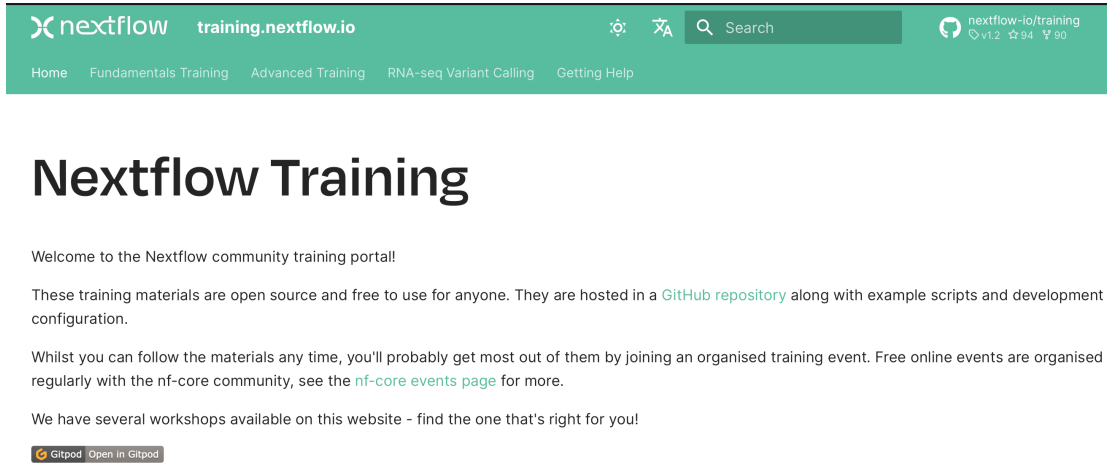
Comparisons were made between sample groups defined using metadata columns, as described in the following table of contrasts:

Table of contrasts

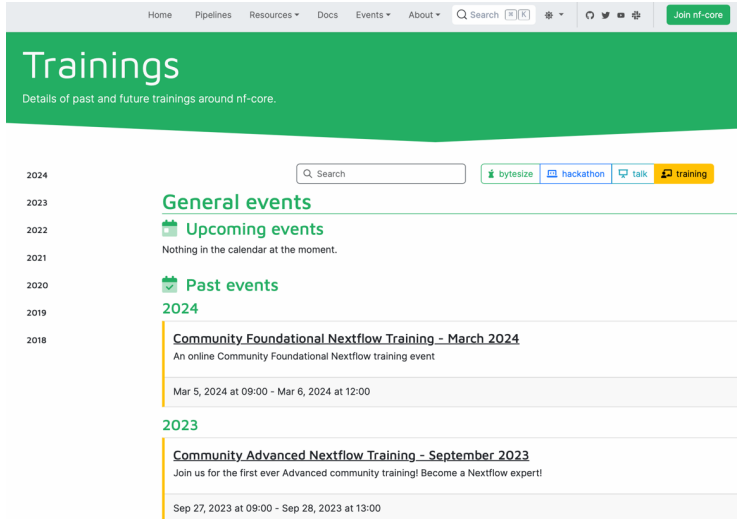
[2024 Spring nextflow and nf-core workshop at Tufts]

What will not be covered

1. Create your own pipelines with nextflow.



https://training.nextflow.io



https://nf-co.re/events/training

2. The concepts of RNA-Seq

https://huoww07.github.io/Bioinformatics-for-RNA-Seq/

Disclaimer

- Some contents are from nextflow/nf-core training materials
- Only my own and Shirley's limited usage

Tufts UNIVERSITY

Day 1
Intro to nextflow & nf-core

Shirley Li
Bioinformatician
TTS Research Technology

Yucheng Zhang
Bioinformatics Engineer
TTS Research Technology

nextflow nf-core  OPEN OnDemand

Overview

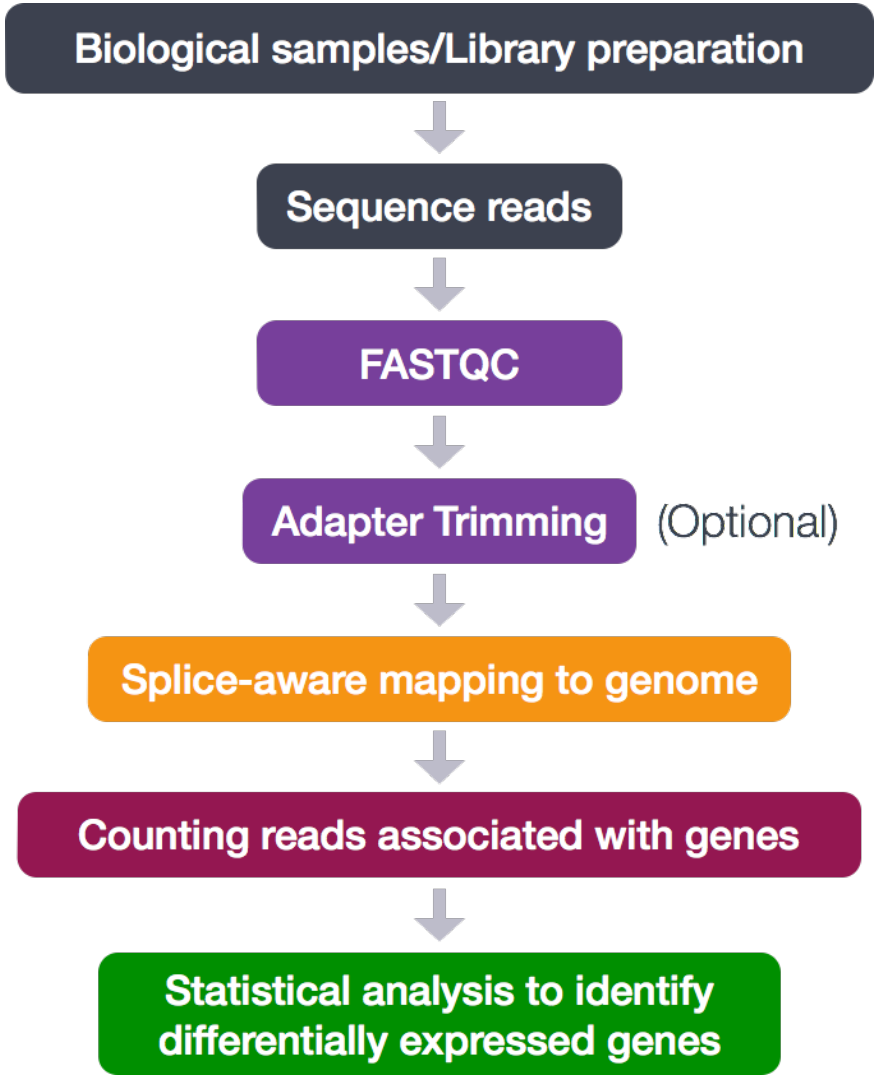
- Nextflow as a Workflow Management Solution
- The nf-core Community and configuring nf-core Pipelines
- Run nf-core pipeline at Tufts HPC
- nf-core ‘fetchngs’ pipeline
- Hands-on demo

Intro to nextflow and nf-core

Nextflow

Workflow

- A workflow is a collection of several analysis steps
- Steps are linked by input/output files
- One often needs to run the same workflow for several samples



Bad workflows

```
## fastp
fastp -i SRR1553607_1.fastq -o SRR1553607_1.fastq.trimmed.fq --max_len1 20
fastp -i SRR1553607_2.fastq -o SRR1553607_2.fastq.trimmed.fq --max_len1 20
fastp -i SRR1972917_1.fastq -o SRR1972917_1.fastq.trimmed.fq --max_len1 20
fastp -i SRR1972917_2.fastq -o SRR1972917_2.fastq.trimmed.fq --max_len1 20
## fastqc
fastqc SRR1553607_1.fastq.trimmed.fq
fastqc SRR1553607_2.fastq.trimmed.fq
fastqc SRR1972917_1.fastq.trimmed.fq
fastqc SRR1972917_1.fastq.trimmed.fq
```

- **Error-prone**
- **Not reuseable**
- **No parallelism**

Better but still bad workflows: for loop

```
## fastp
for name in *.fastq; do
    fastp -i $name -o ${name%.*}.trimmed.fq --max_len1 20
done

## fastqc
for name in *.trimmed.fq; do
    fastpc -i $name
done
```

- For loop runs only one command at a time.
- Our computers have many cores so that we could be run multiple commands at the same time.
- We could add **&** operator to the end of the command to run it in the background.
- But then it runs all commands simultaneously, which we don't want either.
- **We want to run as many commands as we have compute cores, but no more.**

Biostars
BIOINFORMATICS EXPLAINED

Community Planet New Post

Tool:Gnu Parallel - Parallelize Serial Command Line Programs Without Changing Them

293

Article describing tool (for citations):

O. Tange (2011): GNU Parallel - The Command-Line Power Tool, ;login: The USENIX Magazine, February 2011:42-47.

Author's website for obtaining code:

http://www.gnu.org/software/parallel/

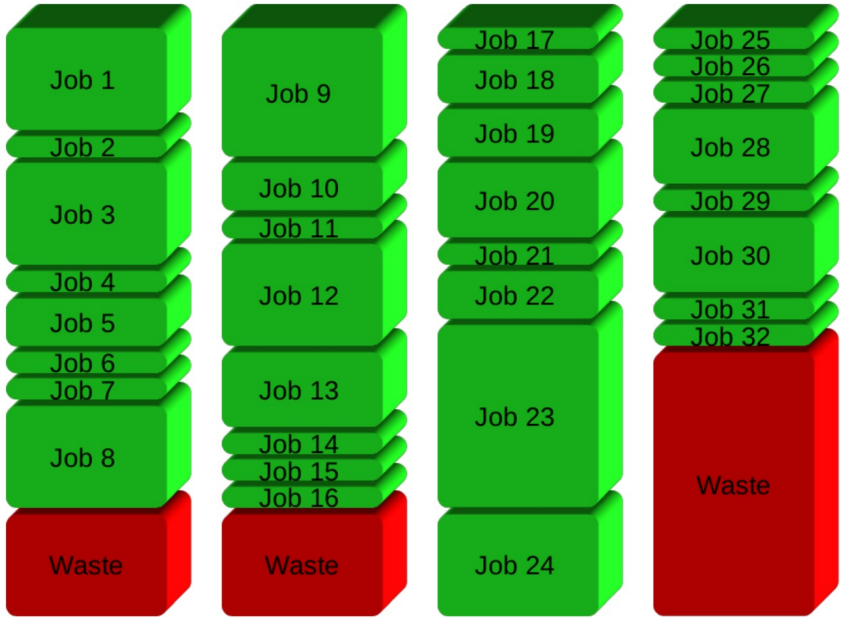
11.1 years ago
ole.tange ★ 4.4k

All new computers have multiple cores. Many bioinformatics tools are serial in nature and will therefore not use the multiple cores. However, many bioinformatics tasks (especially within NGS) are extremely parallelizeable:

- Run the same program on many files
- Run the same program on every sequence

GNU Parallel is a general parallelizer and makes is easy to run jobs in parallel on the same machine or on multiple machines you have ssh access to.

If you have 32 different jobs you want to run on 4 CPUs, a straight forward way to parallelize is to run 8 jobs on each CPU:



https://www.biostars.org/p/63816/

Feature	Nextflow	Snakemake
Language	Groovy (DSL)	Python (DSL)
Dependency Mgmt	Built-in (via Conda, Docker, Singularity, etc.)	Built-in (via Conda, Docker, Singularity, etc.)
Parallelism	Native parallelism	Native parallelism
Syntax	Declarative, flow-based	Declarative, rule-based
Directed Acyclic	Yes	Yes
Ease of Learning	Requires some familiarity with Groovy	Familiarity with Python
Community	**nf-core**	
File Handling	Supports file system operations	Supports file system operations
Error Handling	Robust error handling mechanism	Error handling through Python
Visualization	Built-in visualization tools	External visualization tools
Integration	Strong integration with AWS, GCP, Azure, etc.	Can be integrated with cloud platforms

nextflow *pipeline*

Write code
in any language

Orchestrate tasks with
dataflow programming

Define software
dependencies
via containers

Built-in version
control with Git

nextflow *runtime*

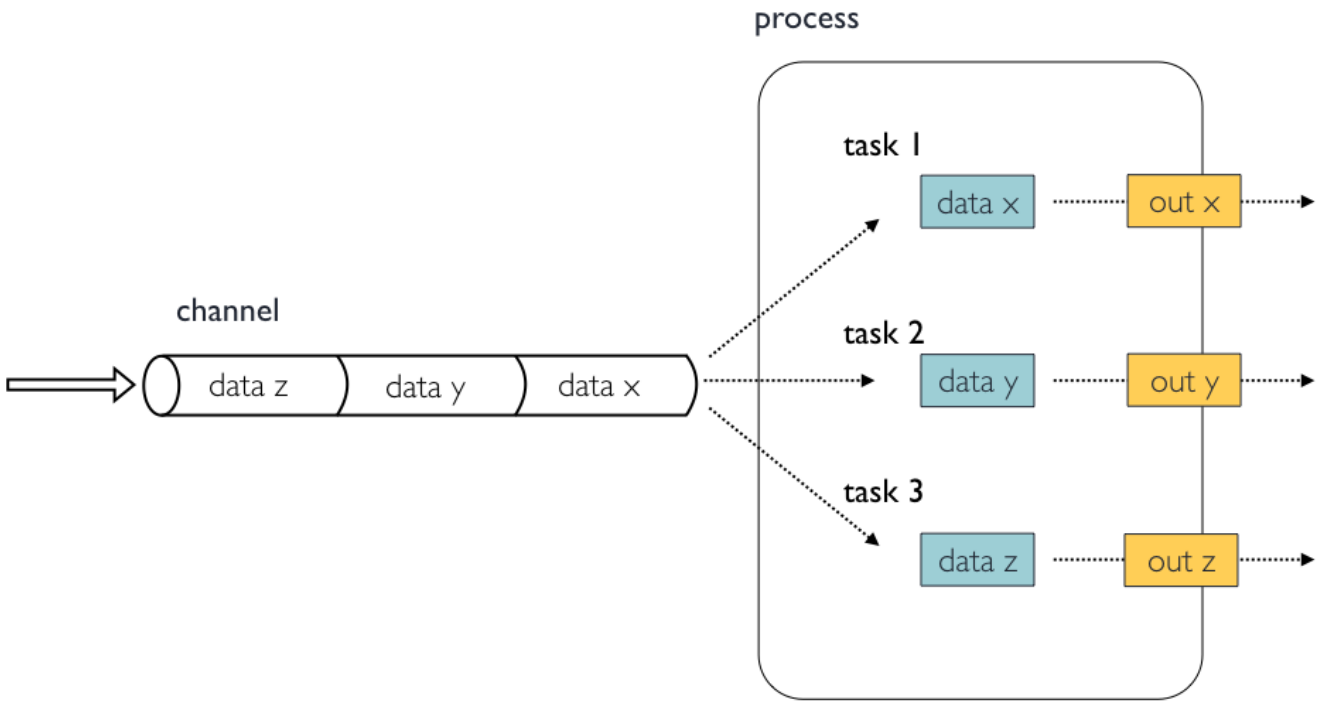
Task orchestration
and execution

Supported Platforms

Building blocks

- **channel**: information flows from one process to another via ‘channels’ as defined in the input and output sections of each process.

- **process**: one (independent) step in the pipeline block. This is where the execution of code happens



The ability to easily parallelize tasks

Example nextflow workflow

Groovy programming languages

```
// Declare syntax version
nextflow.enable.dsl=2

// Script parameters
params.query = "/some/data/sample.fa"
params.db = "/some/path/pdb"

process blastSearch {
  input:
    path query
    path db
  output:
    path "top_hits.txt"

    """
    blastp -db $db -query $query -outfmt 6 > blast_result
    cat blast_result | head -n 10 | cut -f 2 > top_hits.txt
    """
}

process extractTopHits {
  input:
    path top_hits

  output:
    path "sequences.txt"

    """
    blastdbcmd -db $db -entry_batch $top_hits > sequences.txt
    """
}

workflow {
  def query_ch = Channel.fromPath(params.query)
  blastSearch(query_ch, params.db) | extractTopHits | view
}
```

Run a local pipeline

```
#!/usr/bin/env nextflow

params.greeting = 'Hello world!'
greeting_ch = Channel.of(params.greeting)

process SPLITLETTERS {
    input:
    val x

    output:
    path 'chunk_*'

    script:
    """
    printf '$x' | split -b 6 - chunk_
    """
}

process CONVERTTOUPPER {
    input:
    path y

    output:
    stdout

    script:
    """
    cat $y | tr '[a-z]' '[A-Z]'
    """
}

workflow {
    letters_ch = SPLITLETTERS(greeting_ch)
    results_ch = CONVERTTOUPPER(letters_ch.flatten())
    results_ch.view{ it }
}
```

```
[yzhang85@c1cmp063 nf-training]$ nextflow run hello.nf
Nextflow 23.10.1 is available - Please consider updating your version to it
N E X T F L O W  ~  version 23.10.0
Launching `hello.nf` [furious_newton] DSL2 - revision: 3c3d5e1897
executor >  local (3)
[8f/3b8107] process > SPLITLETTERS (1)   [100%] 1 of 1 ✔
[d3/4546d4] process > CONVERTTOUPPER (1) [100%] 2 of 2 ✔
WORLD!
HELLO
```

Run a remote pipeline


```
[tutln02@p1cmp045 test]$ module load nextflow/23.10.0
[tutln02@p1cmp045 test]$ module load singularity
[tutln02@p1cmp045 test]$ nextflow run nf-core/funcscan -r 1.1.5 -profile test,singularity --outdir testout
CAPSULE: Downloading dependency org.multiverse:multiverse-core:jar:0.7.0
CAPSULE: Downloading dependency org.apache.ivy:ivy:jar:2.5.1
CAPSULE: Downloading dependency org.codehaus.groovy:groovy:jar:3.0.19
CAPSULE: Downloading dependency commons-io:commons-io:jar:2.11.0
CAPSULE: Downloading dependency com.google.guava:listenablefuture:jar:9999.0-empty-to-avoid-conflict-with-guava
CAPSULE: Downloading dependency io.nextflow:nf-httpfs:jar:23.10.0
CAPSULE: Downloading dependency com.beust:jcommander:jar:1.35
CAPSULE: Downloading dependency io.nextflow:nf-commons:jar:23.10.0
CAPSULE: Downloading dependency jline:jline:jar:2.9
CAPSULE: Downloading dependency org.codehaus.groovy:groovy-xml:jar:3.0.19
CAPSULE: Downloading dependency org.slf4j:log4j-over-slf4j:jar:2.0.7
CAPSULE: Downloading dependency com.github.zafarkhaja:java-semver:jar:0.9.0
CAPSULE: Downloading dependency javax.mail:mail:jar:1.4.7
```

```
Core Nextflow options
  revision                  : 1.1.5
  runName                   : angry_neumann
  containerEngine           : singularity
  launchDir                 : /cluster/home/tutln02/test
  workDir                   : /cluster/home/tutln02/test/work
  projectDir                : /cluster/home/tutln02/.nextflow/assets/nf-core/funcscan
  userName                  : tutln02
  profile                   : test,singularity
  configFiles               :
```

```
Pulling Singularity image https://depot.galaxyproject.org/singularity/amplify:1.1.0--hdfd78af_0 [cache /cluster/home/tutln02/test/work/s
ingularity/depot.galaxyproject.org-singularity-amplify-1.1.0--hdfd78af_0.img]
Pulling Singularity image https://depot.galaxyproject.org/singularity/hmmer:3.3.2--h1b792b2_1 [cache /cluster/home/tutln02/test/work/sin
gularity/depot.galaxyproject.org-singularity-hmmer-3.3.2--h1b792b2_1.img]
WARN: Singularity cache directory has not been defined -- Remote image will be stored in the path: /cluster/home/tutln02/test/work/singu
larity -- Use the environment variable NXF_SINGULARITY_CACHEDIR to specify a different location
```

To run a pipeline hosted in a remote code repository like github, you simply need to specify its the owner name and the repository name separated by a / character.

SUMMIT 2024

nextflow SUMMIT

Get ready for SUMMIT events in 2024, in Boston and Barcelona!

nextflow SUMMIT
Boston 2024

IN PERSON

May 21-24, 2024

nextflow SUMMIT
Barcelona 2024

IN PERSON | ONLINE

October 28 - November 1, 2024

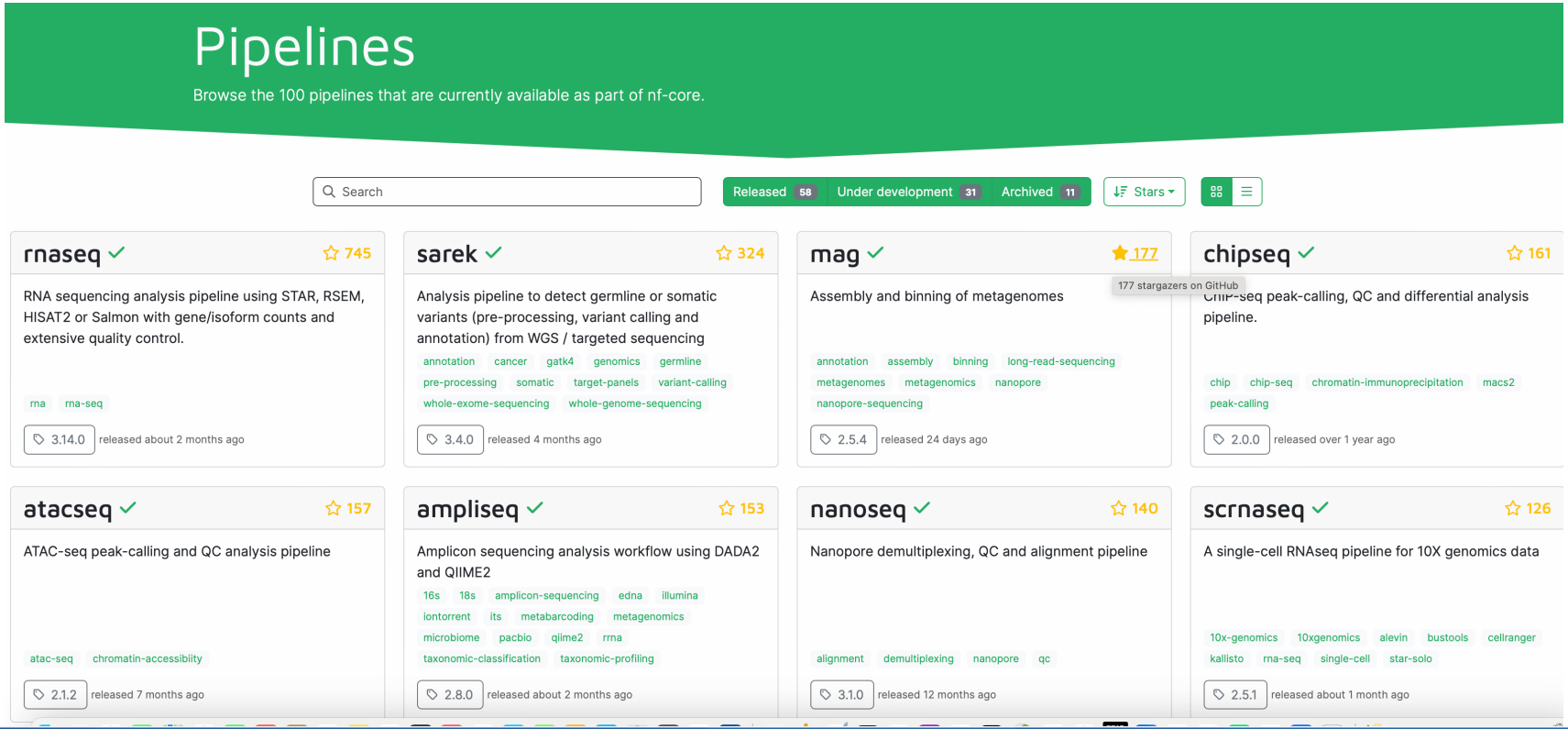
Intro to nextflow and nf-core

nf-core

nf-core

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

https://nf-co.re/pipelines



nf-core tools

```
[yzhang85@c1cmp038 ~]$ module load nf-core
[yzhang85@c1cmp038 ~]$ nf-core --help
```

```
nf-core/tools version 2.13.1 - https://nf-co.re


Usage: nf-core [OPTIONS] COMMAND [ARGS]...

nf-core/tools provides a set of helper tools for use with nf-core Nextflow pipelines.
It is designed for both end-users running pipelines and also developers creating new pipelines.

╭─ Options ──────────────────────────────────────────────────────────────────────────────────────╮
│ --version                        Show the version and exit.                                    │
│ --verbose         -v             Print verbose output to the console.                          │
│ --hide-progress                  Don't show progress bars.                                     │
│ --log-file        -l  <filename> Save a verbose log to a file.                                 │
│ --help            -h             Show this message and exit.                                   │
╰────────────────────────────────────────────────────────────────────────────────────────────────╯
╭─ Commands for users ───────────────────────────────────────────────────────────────────────────╮
│ list                 List available nf-core pipelines with local info.                         │
│ launch               Launch a pipeline using a web GUI or command line prompts.               │
│ create-params-file   Build a parameter file for a pipeline.                                    │
│ download             Download a pipeline, nf-core/configs and pipeline singularity images.     │
│ licences             List software licences for a given workflow (DSL1 only).                  │
│ tui                  Open Textual TUI.                                                         │
╰────────────────────────────────────────────────────────────────────────────────────────────────╯
╭─ Commands for developers ──────────────────────────────────────────────────────────────────────╮
│ create           Create a new pipeline using the nf-core template.                             │
│ lint             Check pipeline code against nf-core guidelines.                               │
│ modules          Commands to manage Nextflow DSL2 modules (tool wrappers).                     │
│ subworkflows     Commands to manage Nextflow DSL2 subworkflows (tool wrappers).                │
│ schema           Suite of tools for developers to manage pipeline schema.                      │
│ create-logo      Generate a logo with the nf-core logo template.                               │
│ bump-version     Update nf-core pipeline version number.                                       │
│ sync             Sync a pipeline TEMPLATE branch with the nf-core template.                    │
╰────────────────────────────────────────────────────────────────────────────────────────────────╯
```

nf-core list

```
[yzhang85@c1cmp038 projects]$ nf-core list --sort stars
```

```
nf-core/tools version 2.13.1 - https://nf-co.re
```

Pipeline Name	Stars	Latest Release	Released	Last Pulled	Have latest release?
rnaseq	755	3.14.0	2 months ago	-	-
sarek	330	3.4.0	4 months ago	-	-
mag	179	2.5.4	1 months ago	-	-
chipseq	164	2.0.0	1 years ago	-	-
scrnaseq	162	2.5.1	2 months ago	-	-
atacseq	157	2.1.2	7 months ago	-	-
ampliseq	155	2.8.0	2 months ago	-	-
nanoseq	143	3.1.0	1 years ago	-	-
methylseq	127	2.6.0	2 months ago	-	-
rnafusion	126	3.0.1	4 months ago	-	-
eager	122	2.5.1	4 weeks ago	-	-
fetchngs	118	1.12.0	3 weeks ago	-	-
viralrecon	105	2.6.0	12 months ago	-	-
taxprofiler	86	1.1.5	1 months ago	-	-
hic	72	2.1.0	10 months ago	-	-
raredisease	67	2.0.0	7 hours ago	-	-
smrnaseq	64	2.3.0	3 weeks ago	-	-
cutandrun	62	3.2.2	2 months ago	-	-
funcscan	52	1.1.4	4 months ago	-	-
bacass	50	2.1.0	5 months ago	-	-
hlatyping	49	2.0.0	1 years ago	-	-
bactmap	47	1.0.0	3 years ago	-	-
pangenome	44	1.1.1	3 days ago	-	-
airrflow	40	3.2.0	5 months ago	-	-
differentialabundance	39	1.4.0	4 months ago	-	-
spatialtranscriptomics	38	dev	2 months ago	-	-
proteinfold	36	1.0.0	1 years ago	-	-
circrna	35	dev	6 days ago	-	-
demultiplex	32	1.4.1	3 weeks ago	-	-

nf-core download

```
[yzhang85@p1cmp028 nf-core]$ nf-core download --help
```

```
    nf-core/tools version 2.13.1 - https://nf-co.re


 Usage: nf-core download [OPTIONS] <pipeline name>

 Download a pipeline, nf-core/configs and pipeline singularity images.
 Collects all files in a single archive and configures the downloaded workflow to use relative
 paths to the configs and singularity images.

╭─ Options ──────────────────────────────────────────────────────────────────────────────────────╮
| --revision                       -r  TEXT                        Pipeline release to download.
|                                                                  Multiple invocations are possible,
|                                                                  e.g. `-r 1.1 -r 1.2`
| --outdir                         -o  TEXT                        Output directory
| --compress                       -x  [tar.gz|tar.bz2|zip|none]   Archive compression type
| --force                          -f                              Overwrite existing files
| --tower                          -t                              Download for Seqera Platform
|                                                                  (formerly Nextflow Tower)
| --download-configuration         -d                              Include configuration profiles in
|                                                                  download. Not available with
|                                                                  `--tower`
| --container-system               -s  [none|singularity]          Download container images of
|                                                                  required software.
| --container-library              -l  TEXT                        Container registry/library or
|                                                                  mirror to pull images from.
| --container-cache-utilisation    -u  [amend|copy|remote]         Utilise a `singularity.cacheDir`
|                                                                  in the download process, if
|                                                                  applicable.
| --container-cache-index          -i  TEXT                        List of images already available
|                                                                  in a remote
|                                                                  `singularity.cacheDir`.
| --parallel-downloads             -p  INTEGER                     Number of parallel image downloads
| --help                           -h                              Show this message and exit.
╰────────────────────────────────────────────────────────────────────────────────────────────────╯

[yzhang85@p1cmp028 nf-core]$ nf-core download rnaseq -r 3.14.0 --outdir rnaseq -d -s singularity -x none
```

Local nf-core pipelines

HPC system administrators have downloaded popular nf-core pipelines and stored them in the following directory:

/cluster/tufts/biocontainers/nf-core/pipelines/

```
[yzhang85@login-prod-03 ~]$ ls /cluster/tufts/biocontainers/nf-core/pipelines/
nf-core-ampliseq/                nf-core-mag/           nf-core-rnasplice/
nf-core-atacseq/                 nf-core-metatdenovo/   nf-core-sarek/
nf-core-chipseq/                 nf-core-methylseq/     nf-core-scrnaseq/
nf-core-differentialabundance/   nf-core-nanoseq/       nf-core-smrnaseq/
nf-core-eager/                   nf-core-nanostring/    nf-core-taxprofiler/
nf-core-fetchngs/                nf-core-pangenome/     nf-core-viralrecon/
nf-core-funcscan/                nf-core-rnafusion/
nf-core-hic/                     nf-core-rnaseq/
```

Usage instructions and documentation

Each pipeline has its own webpage at https://nf-co.re/<pipeline_name> e.g. nf-co.re/rnaseq

```
[yzhang85@login-prod-03 ~]$ nextflow run /cluster/tufts/biocontainers/nf-core/pipelines/nf-core-rnaseq/3.14.0/3_14_0/ --help
Nextflow 23.10.1 is available - Please consider updating your version to it
N E X T F L O W  ~  version 23.10.0
Launching `/cluster/tufts/biocontainers/nf-core/pipelines/nf-core-rnaseq/3.14.0/3_14_0/main.nf` [lonely_wright] DSL2 - revision: 746820de9b


------------------------------------------------------
                                        ,--./,-.
        ___     __   __   __   ___     /,-._.--~'
  |\ | |__  __ /  ` /  \ |__) |__         }  {
  | \| |       \__, \__/ |  \ |___     \`-._,-`-,
                                        `._,._,'

  nf-core/rnaseq v3.14.0
------------------------------------------------------
Typical pipeline command:

  nextflow run nf-core/rnaseq --input samplesheet.csv --genome GRCh37 -profile docker

Input/output options
  --input                       [string]  Path to comma-separated file containing information about the samples in the experiment.
  --outdir                      [string]  The output directory where the results will be saved. You have to use absolute paths to storage
on Cloud
                                          infrastructure.
  --email                       [string]  Email address for completion summary.
  --multiqc_title               [string]  MultiQC report title. Printed as page header, used for filename if not otherwise specified.

Reference genome options
  --genome                      [string]  Name of iGenomes reference.
  --fasta                       [string]  Path to FASTA genome file.
  --gtf                         [string]  Path to GTF annotation file.
  --gff                         [string]  Path to GFF3 annotation file.
```

NXF_SINGULARITY_CACHEDIR

Storing singularity images on our cluster can significantly speed up pipeline execution times, as the images are downloaded once and then used when needed.

The location of such cache directory is defined by the environment variable **NXF_SINGULARITY_CACHEDIR**

If you want to run the nf-core pipe lines managed by system admins, please define NXF_SINGULARITY_CACHEDIR like this:

export NXF_SINGULARITY_CACHEDIR=/cluster/tufts/biocontainers/nf-core/singularity-images

However, if you need to run your own pipelines, you have to define **NXF_SINGULARITY_CACHEDIR** to your own directory. **Please do not use your $HOME**.

Config files

nf-core pipelines make use of nextflow's configuration files to specify how the pipelines runs, define custom parameters and what software management system to use e.g. docker, singularity or conda.

Default base config

```
nextflow run nf-core/<pipeline>
```


Automatically loaded


Sensible default resource allocation


No software packaging specified


Runs locally, no job submission

Core profiles

```
nextflow run nf-core/<pipeline> -profile singularity
```

Specify software packaging

- Docker
- Singularity
- Conda

Specify test profile 

[2024 Spring Containers Workshop at Tufts]

Institutional profiles

```
nextflow run nf-core/<pipeline> -profile mycluster
```

 Specifies job submission

 Specify software packaging

Works for:

 For all pipelines

 For all users on your system

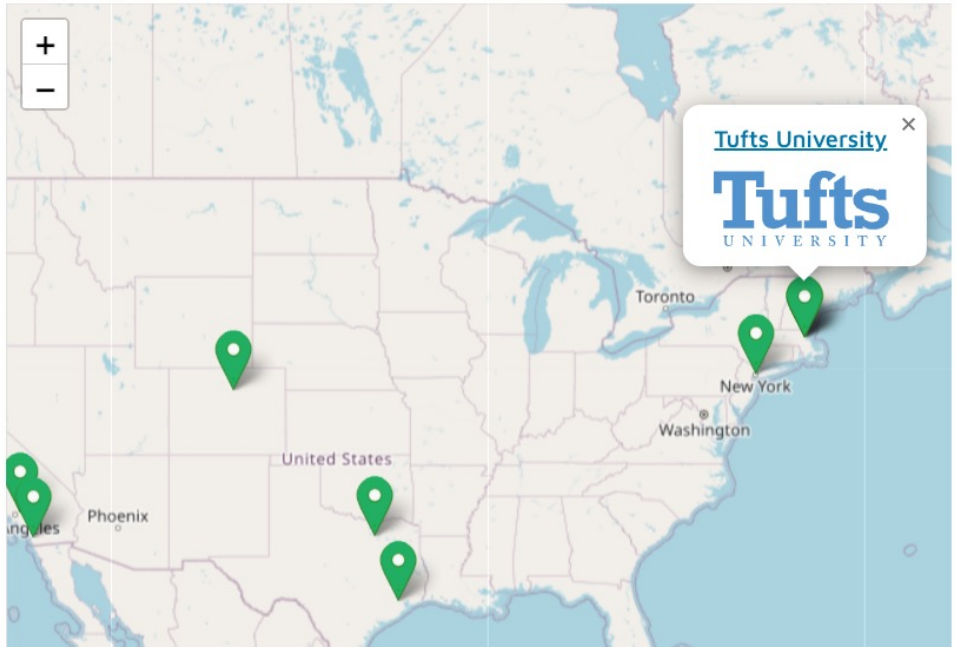
 Single point to update

Tufts as one of the four contributors on the East Coast

Organisations

Some of the organisations running nf-core pipelines are listed below, along with a key person who you can contact for advice.

Note Expand



Tufts University

Tufts University is a prestigious private research university located in Medford, Massachusetts, just outside of Boston. Established in 1852, it has a rich history and a reputation for academic excellence, innovative research, and a commitment to active citizenship and social responsibility.

Yucheng Zhang zhan4429 TuftsUniversity

https://nf-co.re/contributors

tufts profile


```
params {
        max_memory = 120.GB
        max_cpus = 72
        max_time = 168.h
        igenomes_base = '/cluster/tufts/biocontainers/datasets/igenomes/'
}

process {
        executor = 'slurm'
        clusterOptions = '-N 1 -n 1 -p batch'
 }

executor {
        queueSize = 16
        pollInterval = '1 min'
        queueStatInterval = '5 min'
        submitRateLimit = '10 sec'
}

// Set $NXF_SINGULARITY_CACHEDIR in your ~/.bashrc
// to stop downloading the same image for every run
singularity {
        enabled = true
        autoMounts = true
}
```

https://github.com/nf-core/configs/blob/master/conf/tufts.config

test profile

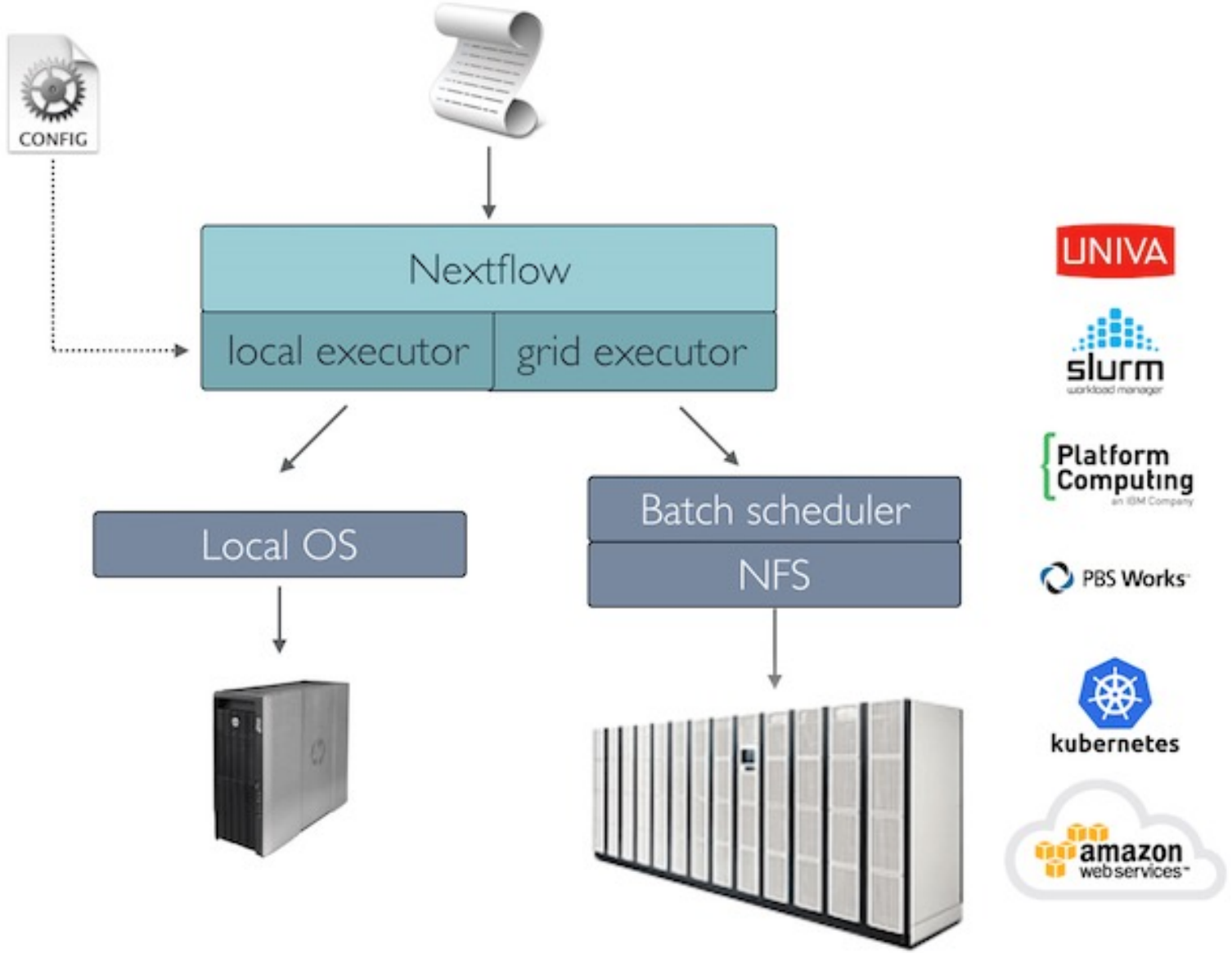
```
[yzhang85@p1cmp044 nf-core_workshop]$ rnaseq -profile test,tufts --outdir output
Nextflow 23.10.1 is available - Please consider updating your version to it
N E X T F L O W  ~  version 23.10.0
Launching `/cluster/tufts/biocontainers/nf-core/pipelines/nf-core-rnaseq/3.14.0/3_14_0/main.nf` [naughty_pesquet] DSL2 - revision: 746820
de9b
WARN: The following invalid input values have been detected:

* --config_profile_contact_github: @zhan4429
* --config_profile_contact_email: Yucheng.Zhang@tufts.edu
* --partition: batch
```

Institutional config options

```
  config_profile_name       : Test profile
  config_profile_description: Minimal test dataset to check pipeline function
  config_profile_contact    : Yucheng Zhang
  config_profile_url        : https://it.tufts.edu/high-performance-computing
```

Intro to nextflow and nf-core

Run nf-core pipelines at Tuft HPC


CONFIG
Nextflow
local executor
grid executor
Local OS
Batch scheduler
NFS
UNIVA
slurm
workload manager
Platform Computing
an IBM Company
PBS Works
kubernetes
amazon web services

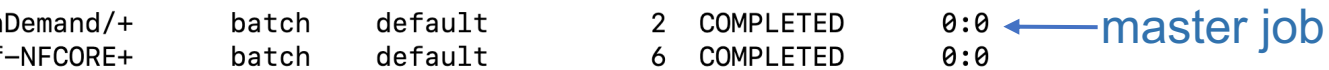
Local mode

```
#!/bin/bash

#SBATCH --time=00-48:00:00
#SBATCH -p batch
#SBATCH -N 1
#SBATCH -n 1
#SBATCH -c XX
#SBATCH --mem=XXG
#SBATCH --job-name nf-core
#SBATCH --output=%x-%J-%u.out
#SBATCH --error=%x-%J-%u.err
#SBATCH --mail-type=ALL
#SBATCH --mail-user=XXX@tufts.edu

module load nf-core
export NXF_SINGULARITY_CACHEDIR=/cluster/tufts/biocontainers/nf-core/singularity-images

nextflow run /cluster/tufts/biocontainers/nf-core/pipelines/nf-core-rnaseq/3.14.0/3_14_0/ \
        --input samplesheet.csv --outdir output \
        --fasta ref.fasta --gtf ref.gtf --aligner star_salmon \
        -profile singularity \
        --max_memory XXG --max_cpus XX
```

Tufts profile

```
#!/bin/bash

#SBATCH --time=00-48:00:00
#SBATCH -p batch
#SBATCH -N 1
#SBATCH -n 1
#SBATCH -c 2 ## This is the parent script used for submitting children slurm jobs, 2 cores are enough
#SBATCH --job-name nf-core
#SBATCH --output=%x-%J-%u.out
#SBATCH --error=%x-%J-%u.err
#SBATCH --mail-type=ALL
#SBATCH --mail-user=XXX@tufts.edu

module load nf-core

export NXF_SINGULARITY_CACHEDIR=/cluster/tufts/biocontainers/nf-core/singularity-images

nextflow run /cluster/tufts/biocontainers/nf-core/pipelines/nf-core-rnaseq/3.14.0/3_14_0/ \
        --input samplesheet.csv --outdir output \
        --fasta ref.fasta --gtf ref.gtf \
        --aligner star_salmon \
        -profile tufts
```

2578439	OnDemand/+	batch	default	2	COMPLETED	0:0
2578451	nf-NFCORE+	batch	default	6	COMPLETED	0:0
2578452	nf-NFCORE+	batch	default	12	COMPLETED	0:0
2578453	nf-NFCORE+	batch	default	6	COMPLETED	0:0
2578454	nf-NFCORE+	batch	default	12	COMPLETED	0:0
2578455	nf-NFCORE+	batch	default	6	COMPLETED	0:0
2578456	nf-NFCORE+	batch	default	12	COMPLETED	0:0
2578457	nf-NFCORE+	batch	default	6	COMPLETED	0:0
2578458	nf-NFCORE+	batch	default	12	COMPLETED	0:0
2578459	nf-NFCORE+	batch	default	12	COMPLETED	0:0
2578460	nf-NFCORE+	batch	default	6	COMPLETED	0:0
2578461	nf-NFCORE+	batch	default	6	COMPLETED	0:0
2578462	nf-NFCORE+	batch	default	12	COMPLETED	0:0
2578477	nf-NFCORE+	batch	default	1	COMPLETED	0:0
2578630	nf-NFCORE+	batch	default	1	COMPLETED	0:0
2578692	nf-NFCORE+	batch	default	1	COMPLETED	0:0
2578693	nf-NFCORE+	batch	default	1	COMPLETED	0:0
2578710	nf-NFCORE+	batch	default	12	COMPLETED	0:0
2578711	nf-NFCORE+	batch	default	12	COMPLETED	0:0
2578712	nf-NFCORE+	batch	default	2	COMPLETED	0:0
2578786	nf-NFCORE+	batch	default	1	COMPLETED	0:0
2578909	nf-NFCORE+	batch	default	1	COMPLETED	0:0
2579166	nf-NFCORE+	batch	default	1	COMPLETED	0:0
2579310	nf-NFCORE+	batch	default	1	COMPLETED	0:0
2580524	nf-NFCORE+	batch	default	1	COMPLETED	0:0
2580697	nf-NFCORE+	batch	default	1	COMPLETED	0:0
2580704	nf-NFCORE+	batch	default	6	COMPLETED	0:0
2583415	nf-NFCORE+	batch	default	6	COMPLETED	0:0
2583416	nf-NFCORE+	batch	default	6	COMPLETED	0:0
2583417	nf-NFCORE+	batch	default	6	COMPLETED	0:0
2583418	nf-NFCORE+	batch	default	6	COMPLETED	0:0
2583421	nf-NFCORE+	batch	default	6	COMPLETED	0:0
2583422	nf-NFCORE+	batch	default	6	COMPLETED	0:0
2583439	nf-NFCORE+	batch	default	12	COMPLETED	0:0
2583453	nf-NFCORE+	batch	default	12	COMPLETED	0:0
2583470	nf-NFCORE+	batch	default	12	COMPLETED	0:0
2583483	nf-NFCORE+	batch	default	12	COMPLETED	0:0



Other partitions

```
#!/bin/bash

#SBATCH --time=00-48:00:00
#SBATCH -p batch
#SBATCH -N 1
#SBATCH -n 2   ## This is the parent script used for submitting children slurm jobs, 2 cores are enough
#SBATCH --job-name nf-core
#SBATCH --output=%x-%J-%u.out
#SBATCH --error=%x-%J-%u.err
#SBATCH --mail-type=ALL
#SBATCH --mail-user=XXX@tufts.edu

module load nf-core

export NXF_SINGULARITY_CACHEDIR=/cluster/tufts/biocontainers/nf-core/singularity-images

nextflow run /cluster/tufts/biocontainers/nf-core/pipelines/nf-core-rnaseq/3.14.0/3_14_0/ \
        --input samplesheet.csv --outdir output \
        --fasta ref.fasta --gtf ref.gtf \
        --aligner star_salmon \
        -profile tufts --partition preempt
```

nf-core pipelines as modules

```
[yzhang85@login-prod-03 ~]$ module avail nf-core

---------------------- /cluster/tufts/hpc/tools/module -----------
nf-core/2.10   nf-core/2.13.1

--------------------- /cluster/tufts/biocontainers/modules --------
nf-core-ampliseq/2.8.0              nf-core-nanoseq/3.1.0
nf-core-atacseq/2.1.2               nf-core-nanostring/1.2.1
nf-core-chipseq/2.0.0               nf-core-pangenome/1.1.0
nf-core-differentialabundance/1.4.0 nf-core-pangenome/1.1.1
nf-core-eager/2.5.1                 nf-core-rnafusion/3.0.1
nf-core-fetchngs/1.11.0             nf-core-rnaseq/3.14.0
nf-core-fetchngs/1.12.0             nf-core-rnasplice/1.0.2
nf-core-funcscan/1.1.4              nf-core-rnasplice/1.0.3
nf-core-hic/2.1.0                   nf-core-sarek/3.4.0
nf-core-mag/2.5.2                   nf-core-scrnaseq/2.5.1
nf-core-mag/2.5.4                   nf-core-smrnaseq/2.3.0
nf-core-metatdenovo/1.0.0           nf-core-taxprofiler/1.1.5
nf-core-methylseq/2.6.0             nf-core-viralrecon/2.6.0
```

```
[yzhang85@login-prod-01 ~]$ module show nf-core-rnaseq/3.14.0
-------------------------------------------------------------------
/cluster/tufts/biocontainers/modules/nf-core-rnaseq/3.14.0:

module-whatis    nf-core rnaseq pipeline
module-whatis    https://nf-co.re/rnaseq
prepend-path     PATH /cluster/tufts/biocontainers/tools/nf-core-rnaseq/3.14.0/bin
-------------------------------------------------------------------

[yzhang85@login-prod-01 ~]$ more /cluster/tufts/biocontainers/tools/nf-core-rnaseq/3.14.0/bin/rnaseq
#!/usr/bin/env bash

if [ ! $(command -v singularity) ]; then
        module load singularity
fi

VER=3.14.0
PKG=nf-core-rnaseq

export NXF_SINGULARITY_CACHEDIR=/cluster/tufts/biocontainers/nf-core/singularity-images
nextflow run /cluster/tufts/biocontainers/nf-core/pipelines/nf-core-rnaseq/3.14.0/3_14_0 "$@"
[yzhang85@login-prod-01 ~]$ module load nf-core-rnaseq/3.14.0
[yzhang85@login-prod-01 ~]$ rnaseq --help
Nextflow 23.10.1 is available - Please consider updating your version to it
N E X T F L O W  ~  version 23.10.0
Launching `/cluster/tufts/biocontainers/nf-core/pipelines/nf-core-rnaseq/3.14.0/3_14_0/main.nf` [cranky_hopper] DSL2 - revision: 74
6820de9b


------------------------------------------------------
                                          ,--./,-.
        ___     __   __   __   ___      /,-._.--~'
  |\ | |__  __ /  ` /  \ |__) |__           } {
  | \| |       \__, \__/ |  \ |___     \`-._,-`-,
                                          `-._,-'
```

```
[yzhang85@login-prod-01 ~]$ module show nf-core-chipseq/2.0.0
-------------------------------------------------------------------
/cluster/tufts/biocontainers/modules/nf-core-chipseq/2.0.0:

module-whatis    nf-core chipseq pipeline
module-whatis    https://nf-co.re/chipseq
prepend-path     PATH /cluster/tufts/biocontainers/tools/nf-core-chipseq/2.0.0/bin
-------------------------------------------------------------------

[yzhang85@login-prod-01 ~]$ more /cluster/tufts/biocontainers/tools/nf-core-chipseq/2.0.0/bin/chipseq
#!/usr/bin/env bash

if [ ! $(command -v singularity) ]; then
     module load singularity
fi

VER=2.0.0
PKG=nf-core-chipseq

export NXF_SINGULARITY_CACHEDIR=/cluster/tufts/biocontainers/nf-core/singularity-images
nextflow run /cluster/tufts/biocontainers/nf-core/pipelines/nf-core-chipseq/2.0.0/2_0_0 "$@"
[yzhang85@login-prod-01 ~]$ module load nf-core-chipseq/2.0.0
[yzhang85@login-prod-01 ~]$ chipseq --help
Nextflow 23.10.1 is available - Please consider updating your version to it
N E X T F L O W  ~  version 23.10.0
Launching `/cluster/tufts/biocontainers/nf-core/pipelines/nf-core-chipseq/2.0.0/2_0_0/main.nf` [astonishing_goldberg] DSL2 - revisi
on: 7341307235


------------------------------------------------------
                                        ,--./,-.
                                       /,-._.--~'
  ___     __   __   __   ___            } {
 |\ | |__  __ /  ` /  \ |__) |__         } {
 | \| |       \__, \__/ |  \ |___     \`-._,-`-,
                                       `-._,-'
```

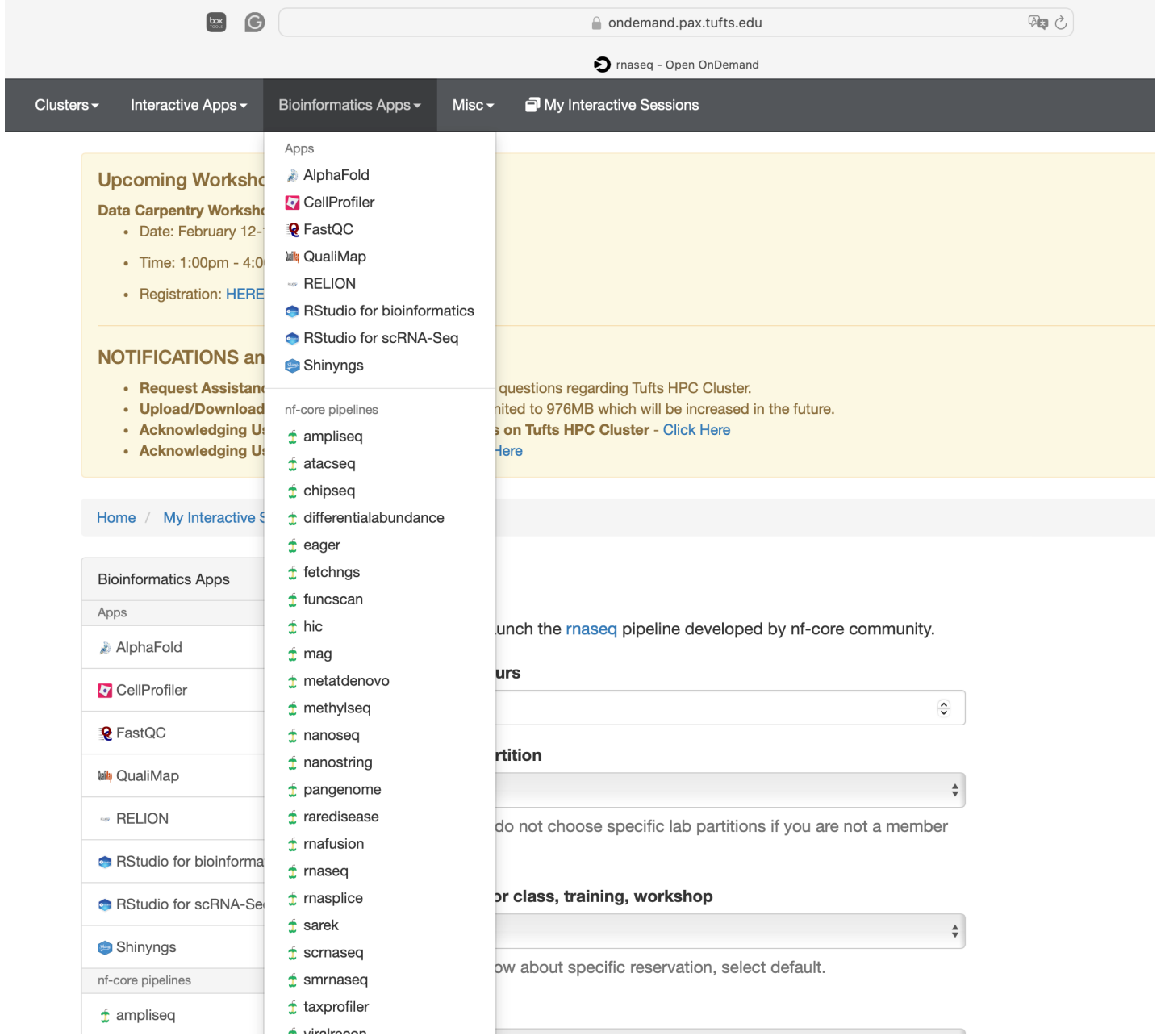
Run pipelines easily with modules

```
#!/bin/bash

#SBATCH --time=00-48:00:00
#SBATCH -p batch
#SBATCH -N 1
#SBATCH -n 1
#SBATCH -c 2 ## This is the parent script used for submitting children slurm jobs, 2 cores are enough
#SBATCH --job-name nf-core
#SBATCH --output=%x-%J-%u.out
#SBATCH --error=%x-%J-%u.err
#SBATCH --mail-type=ALL
#SBATCH --mail-user=XXX@tufts.edu

module load nf-core-rnaseq/3.14.0

rnaseq --input samplesheet.csv --outdir output \
            --fasta ref.fasta --gtf ref.gtf \
            --aligner star_salmon \
            -profile tufts
```

nf-core

+

OPEN OnDemand

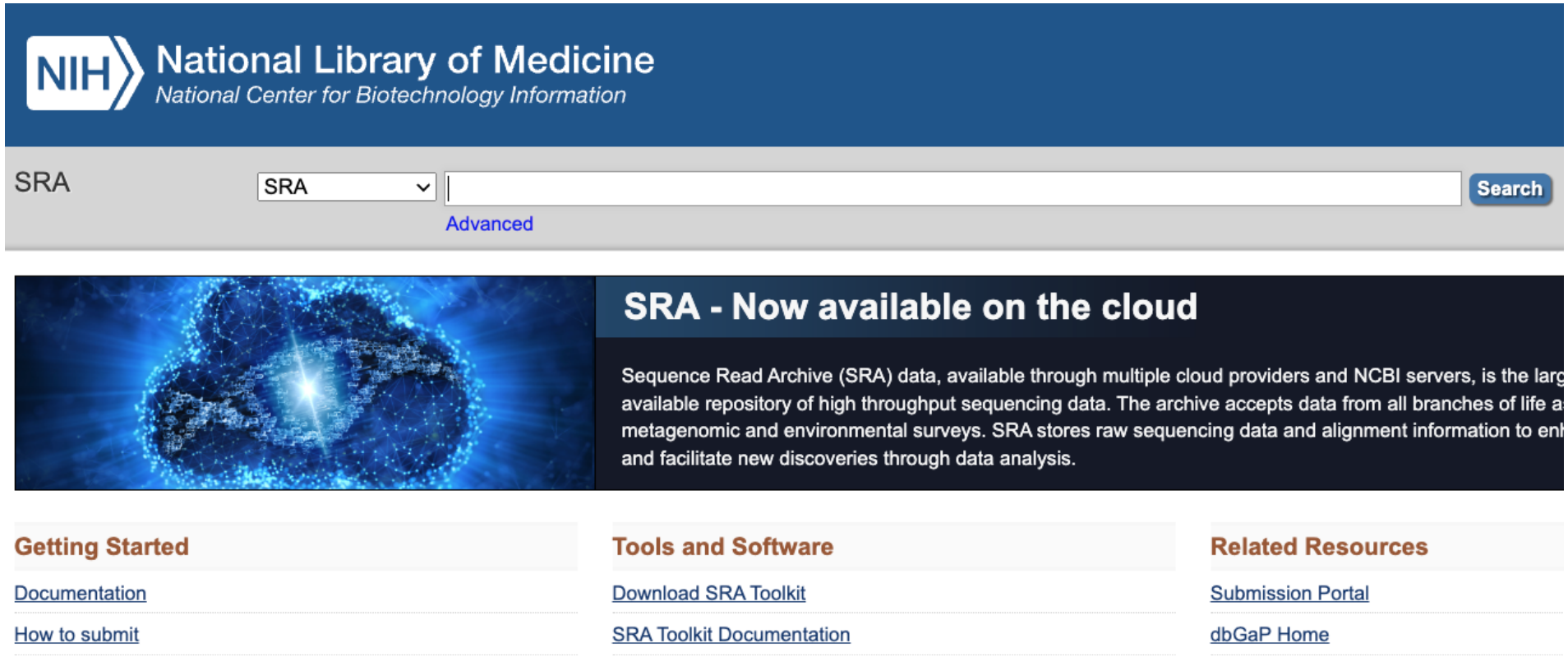


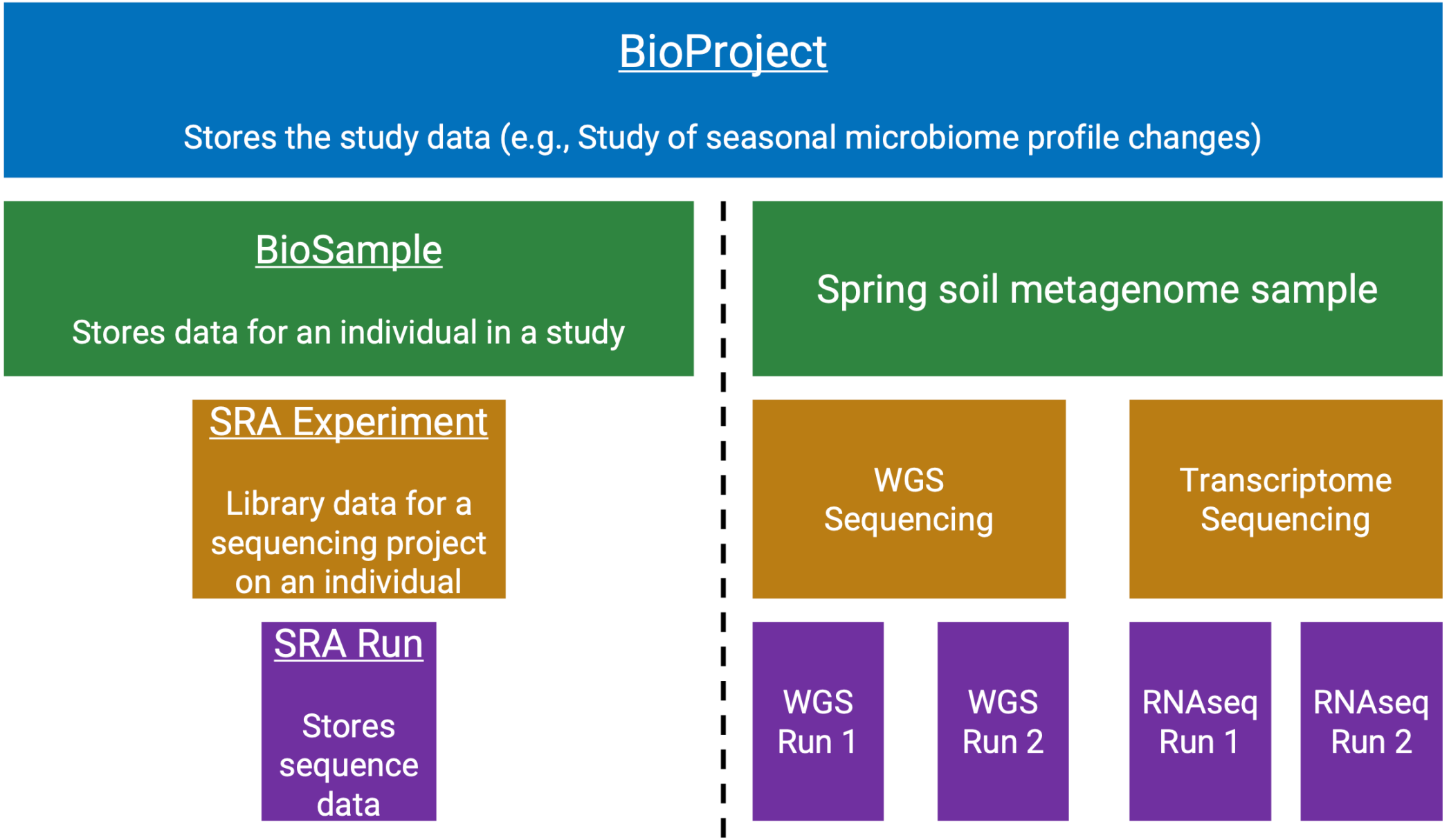
Intro to nextflow and nf-core

nf-core fetchngs

Introduction to SRA (Sequence Read Archive)

- Collection of user-submitted nucleotide sequencing reads, most of which are publicly available to download




BioProject
Stores the study data (e.g., Study of seasonal microbiome profile changes)
BioSample
Stores data for an individual in a study
Spring soil metagenome sample
SRA Experiment
Library data for a sequencing project on an individual
WGS Sequencing
Transcriptome Sequencing
SRA Run
Stores sequence data
WGS Run 1
WGS Run 2
RNAseq Run 1
RNAseq Run 2

Journal of Medicinal Chemistry

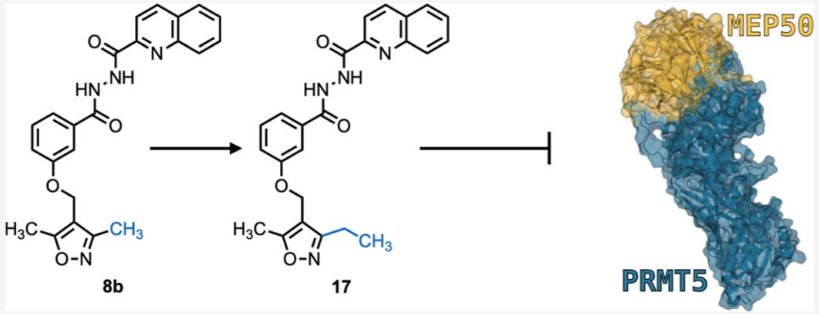
pubs.acs.org/jmc Article

Discovery and Biological Characterization of PRMT5:MEP50 Protein–Protein Interaction Inhibitors

Andrew M. Asberry, Xinpei Cai, Xuehong Deng, Ulises Santiago, Sheng Liu, Hunter S. Sims, Weida Liang, Xueyong Xu, Jun Wan, Wen Jiang, Carlos J. Camacho,* Mingji Dai,* and Chang-Deng Hu*

Cite This: *J. Med. Chem.* 2022, 65, 13793–13812 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Protein arginine methyltransferase 5 (PRMT5) is a master epigenetic regulator and an extensively validated therapeutic target in multiple cancers. Notably, PRMT5 is the only PRMT that requires an obligate cofactor, methylosome protein 50 (MEP50), to function. We developed compound **17**, a novel small-molecule PRMT5:MEP50 protein–protein interaction (PPI) inhibitor, after initial virtual screen hit identification and analogue refinement. Molecular docking indicated that compound **17** targets PRMT5:MEP50 PPI by displacing the MEP50 W54 burial into a hydrophobic pocket of the PRMT5 TIM barrel. *In vitro* analysis indicates IC_{50} < 500 nM for prostate and lung cancer cells with selective, specific inhibition of PRMT5:MEP50 substrate methylation and target gene expression, and RNA-seq analysis suggests that compound **17** may dysregulate TGF-β signaling. Compound **17** provides a proof of concept in targeting PRMT5:MEP50 PPI, as opposed to catalytic targeting, as a novel mechanism of action and supports further preclinical development of inhibitors in this class.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jmedchem.2c01000.

Details of biological reagents (shRNA sequences, antibodies, and oligonucleotide primers) (XLSX)

Compound SMILES and biological data (CSV)

Overall survival curves, PRMT5 and MEP50 expression correlation, PRMT5 and MEP50 electrostatic interactions and binding energy, PRMT5 and MEP50 mutants in BiFC screen, detailed predicted binding mode of Cpd **17**, cryo-EM micrographs of PRMT5/MEP50 and PRMT5/MEP50/Cpd **17**, IC_{50} of Cpd **17** in A549 cancer cell line, and identity/purity data including NMR spectra and HPLC traces (PDF)

Molecular docking structure (PDB)

RNA-seq data for treatment of LNCaP cells with Cpd **17** or shRNA-mediated knockdown of PRMT5 or MEP50 in LCNaP cells are openly available in Gene Expression Omnibus at GSE206460 (Cpd **17** treatment), GSE206820 (MEP50 knockdown), and GSE206111 (PRMT5 knockdown).

RNA-seq data for the knockdown of PRMT5 or MEP50 in A549 cells are openly available in Gene Expression Omnibus at GSE80182.

Series GSE80182

Query DataSets for GSE80182

Status	Public on Jun 09, 2016
Title	A TGFbeta-PRMT5-MEP50 Axis Regulates Cancer Cell Invasion through Histone H3 and H4 Arginine Methylation Coupled Transcriptional Activation and Repression
Organism	Homo sapiens
Experiment type	Expression profiling by high throughput sequencing
Summary	We sequenced mRNA from 3 biological replicates each of A549 lung adenocarcinoma cell lines expressing shRNA against GFP (control), PRMT5, or MEP50. We then determined differential gene expression.
Overall design	Transcriptome analysis of mRNA testing the role of PRMT5 and MEP50 by knockdown in A549 human lung adenocarcinoma cells
Web link	http://www.nature.com/onc/journal/vaop/ncurrent/full/onc2016205a.html
Contributor(s)	Chen H, Shechter D
Citation(s)	Chen H, Lorton B, Gupta V, Shechter D. A TGFβ-PRMT5-MEP50 axis regulates cancer cell invasion through histone H3 and H4 arginine methylation coupled transcriptional activation and repression. *Oncogene* 2017 Jan 19;36(3):373-386. PMID: 27270440

Introduction to GEO (Gene Expression Omnibus)

Platforms (1) GPL16791 Illumina HiSeq 2500 (Homo sapiens)

Samples (9)
Less...

GSM2114336 A549_GFPkd_1
GSM2114337 A549_GFPkd_2
GSM2114338 A549_GFPkd_3
GSM2114339 A549_PRMT5kd_1
GSM2114340 A549_PRMT5kd_2
GSM2114341 A549_PRMT5kd_3
GSM2114342 A549_MEP50kd_1
GSM2114343 A549_MEP50kd_2
GSM2114344 A549_MEP50kd_3

3 samples for control group

3 samples for PRMT5 knockdown group

Relations

BioProject PRJNA318251 ← Bioproject ID

SRA SRP073189 ← SRA runs collection

1. GSM2114344: A549_MEP50kd_3; Homo sapiens; RNA-Seq
   1 ILLUMINA (Illumina HiSeq 2500) run: 38.5M spots, 11.5G bases, 6.7Gb downloads
   Accession: SRX1693959

2. GSM2114343: A549_MEP50kd_2; Homo sapiens; RNA-Seq
   1 ILLUMINA (Illumina HiSeq 2500) run: 32.1M spots, 9.6G bases, 5.6Gb downloads
   Accession: SRX1693958

3. GSM2114342: A549_MEP50kd_1; Homo sapiens; RNA-Seq
   1 ILLUMINA (Illumina HiSeq 2500) run: 30.1M spots, 9G bases, 5.3Gb downloads
   Accession: SRX1693957

4. GSM2114341: A549_PRMT5kd_3; Homo sapiens; RNA-Seq
   1 ILLUMINA (Illumina HiSeq 2500) run: 39.1M spots, 11.7G bases, 6.8Gb downloads
   Accession: SRX1693956

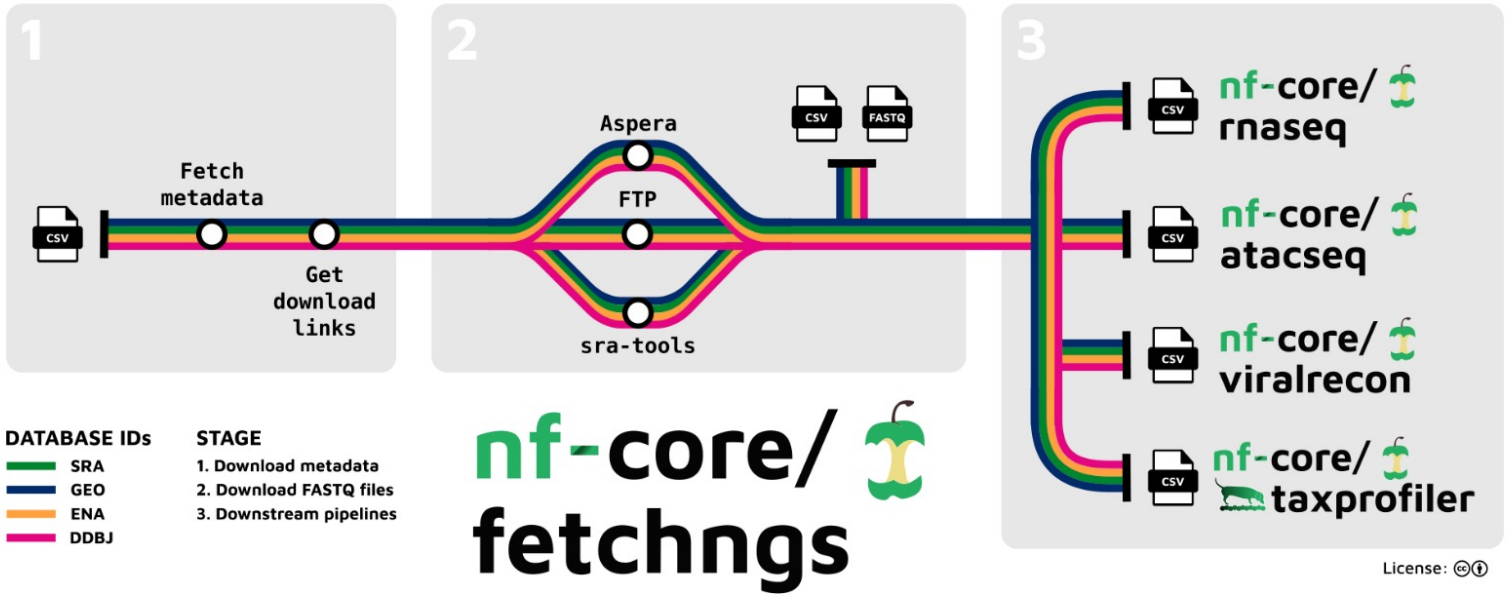
5. GSM2114340: A549_PRMT5kd_2; Homo sapiens; RNA-Seq
   1 ILLUMINA (Illumina HiSeq 2500) run: 40.6M spots, 12.2G bases, 7.1Gb downloads
   Accession: SRX1693955

6. GSM2114339: A549_PRMT5kd_1; Homo sapiens; RNA-Seq
   1 ILLUMINA (Illumina HiSeq 2500) run: 33.9M spots, 10.2G bases, 5.9Gb downloads
   Accession: SRX1693954

Introduction to fetchngs

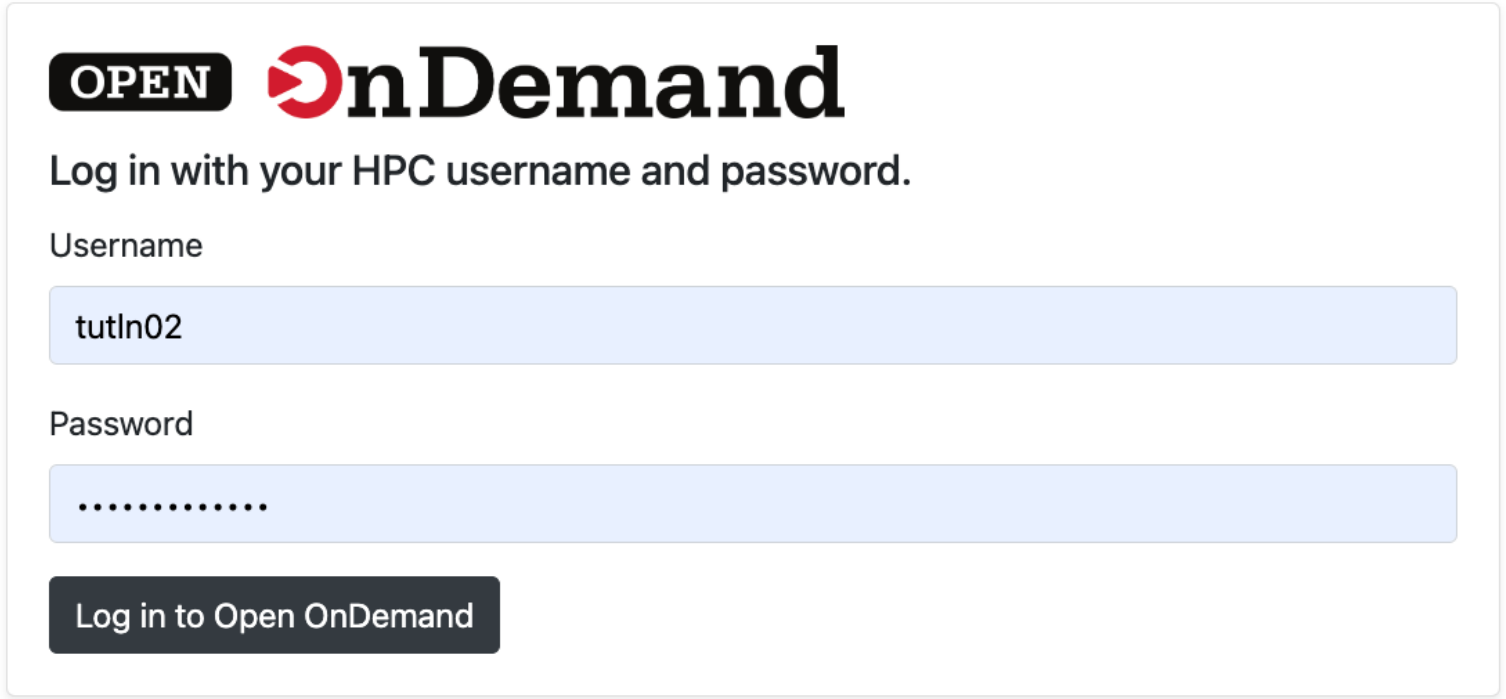
Introduction

nf-core/fetchngs is a bioinformatics pipeline to fetch metadata and raw FastQ files from both public databases. At present, the pipeline supports SRA / ENA / DDBJ / GEO ids (see usage docs).



Run nf-core pipeline with open OnDemand

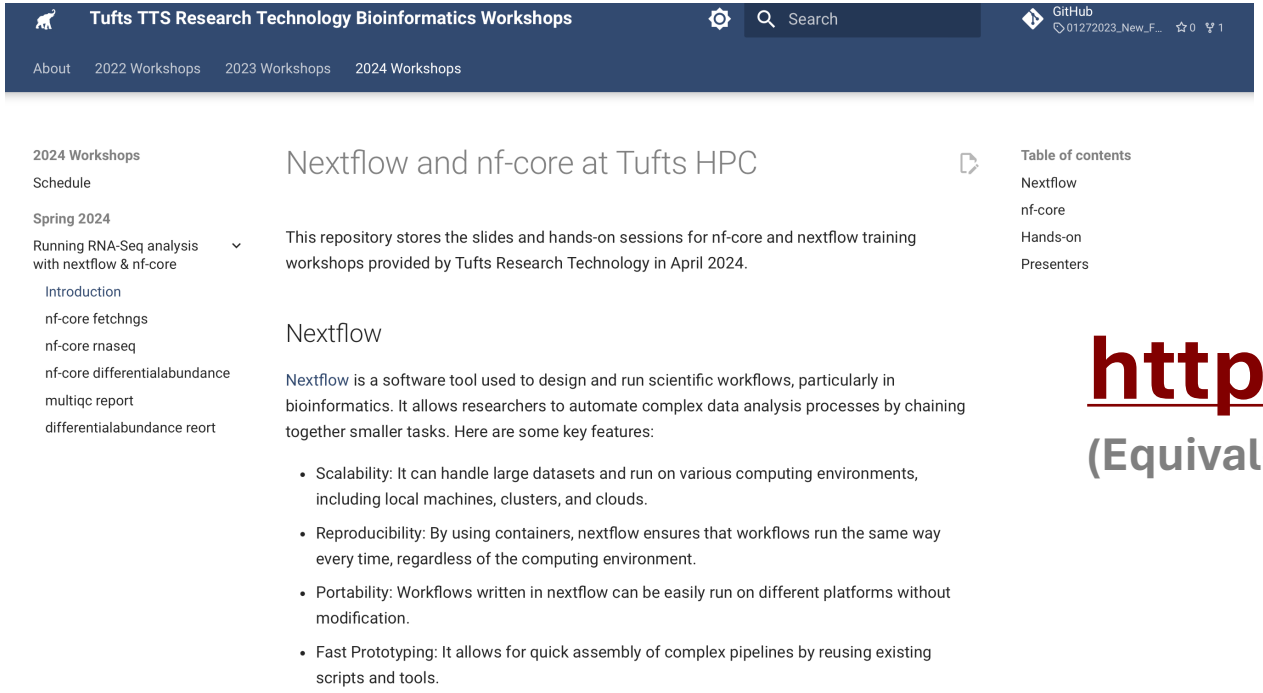
https://ondemand.pax.tufts.edu/



Workshop repo and hands-on

Github repo:
https://github.com/tuftsdatalab/tuftsWorkshops/tree/main/docs/2024_workshops/nfcore_rnaseq_sp24

Webpage:
https://tuftsdatalab.github.io/tuftsWorkshops/2024_workshops/nfcore_rnaseq_sp24/00_introduction/



https://go.tufts.edu/rnaseq

(Equivalent go link, may expire after a few days)